University of Alberta

Population genomics of North American grey wolves (*Canis lupus*)

by

James Christopher Knowles

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science
in
Systematics and Evolution

Biological Sciences

Fall 2010
Edmonton, Alberta

Examining Committee

David Coltman, Biological Sciences

Jocelyn Hall, Biological Sciences

Stephen Moore, Agricultural, Food and Nutritional Science

ABSTRACT

Previous studies of the grey wolf (*Canis lupus*) using microsatellites have showed strong population structure despite the high mobility of individuals. I re-assessed the structure of North American grey wolves by genotyping 132 wolves at a genome-wide set of >26 000 single nucleotide polymorphisms (SNPs), and found less population structure, a strong pattern of isolation by distance, and determined that gene flow between subpopulations relates to prey specialization. To assess how accurately smaller data sets assign individuals, I analyzed sub-sets of SNPs and found that small marker sets varied greatly in estimates of subpopulation assignment, and showed high discordance with assignments determined when using all 26k markers. Finally, using a genome scan to detect natural selection I identified SNPs in three genes that may have undergone directional selection, contain variation with observed phenotypic consequences in other mammal species and may be related to adaptation in grey wolves.

ACKNOWLEDGEMENTS

I am most grateful for the support and guidance of my supervisor, David Coltman, for providing me the opportunity to study in his lab. I truly appreciated his advice and patience during my extended stay at the U of A. Thanks also to my committee advisors Jocelyn Hall and Stephen Moore.

Many people at UCLA were either directly involved in or helped out with my research. Of course, none of the research would have been possible without Bob Wayne, who provided me with funding, a place in his laboratory, and much advice. John Novembre provided practical and theoretical assistance with analysis on many occasions. John Pollinger was helpful during many stages of both laboratory work and data processing. Katie Semple-Delaney instructed me in the use of ARC software, and Eunjung Han gave me access to some necessary data. Last but certainly not least Bridgett vonHoldt was an immense help to me throughout my research. She provided instruction and guidance on everything from laboratory work to data processing and analysis.

I must also thank several people at the University of Alberta who provided assistance along the way. Jocelyn Poissant routinely provided guidance and encouragement, and Aaron Shafer kindly helped with editing and made useful suggestions for the manuscript. Patrick James helped me get my head around various aspects of ordination analyses. Thanks to Bill Clark who aided in tissue sub-sampling while I was away at UCLA. Charlene Nielsen was most helpful with assistance using ARC software.

Tissue samples for this project were made available by Lindsey Carmichael, Marco Musiani, Link Olson, Brandy Jacobsen, Parks Canada, and the University of Alaska Museum. Lindsey and Marco both kindly allowed the use of their genetic and phenotypic data, for which I am most thankful, and Marco in particular provided advice and encouragement during the course of this research.

On a more personal level, I sincerely appreciated the support of all friends, family, and lab-mates. In particular, both my parents, Judith L. Eger and James A. Knowles, provided advice and assistance from the beginning. Finally, I'd like to thank Amber Polywkan for always being there for me during the past three years, even when 'being there' meant moving to Edmonton for the final stage of my research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1
General Introduction

Grey wolf taxonomic history

Grey wolves (*Canis lupus*; Linnaeus, 1758; henceforth "wolves") are large predators, extant across the northern hemisphere, and are recognized as having been the most widely distributed of all wild terrestrial mammals (Nowak 2003; Mech and Boitani 2004b) before extirpation from much of Europe (Boitani 1995), the contiguous United States, and Mexico (Mech 1970). Grey wolves are members of the Canidae (dog) family of the order Carnivora. While the oldest recovered canid fossil dates to around 40 million years ago (mya) during the Eocene (Wang and Tedford 1994), extant canids are thought to share a common ancestor as recently as 12 mya (Wayne *et al.* 1991). Grey wolves are highly derived within the canid phylogeny, and are the most closely related species to the domestic dog, which is hypothesized to have been derived from grey wolves during one or more domestication events (Savolainen *et al.* 2002; Vilà *et al.* 2005; vonHoldt *et al.* 2010) that occurred between 15-100 kya (Vilà *et al.* 1997; Savolainen *et al.* 2002). In fact, the taxonomic status of the dog as a distinct species (*Canis familiaris*; Linnaeus, 1758), rather than a subspecies of grey wolf (*Canis lupus familiaris*; as per Wozencraft 1993), is still under debate. According to a phylogeny of the Canidae generated using ~15k bp of mitochondrial and nuclear sequence (Lindblad-Toh *et al.* 2005), the grey wolf's closest wild relative is the coyote (*Canis latrans*), a smaller North American predator which diverged from the grey wolf in the Lower Pleistocene, ~1–1.5 Mya according to fossil evidence (Nowak 1979).

Extant grey wolves are currently recognized as being composed of 15-20 subspecies across Eurasia and North America (not including the domestic dog; Nowak 2003). Previously, as many as 24 subspecies of grey wolf have been recognized in North America alone (Hall 1981), but this number has since been

reduced to five subspecies by Nowak (1995). Nowak (1995) suggested that separation into different refugia during a previous glacial period led to the extant subspecies. Subspecies recognition in these and other studies has been based on skull morphology distinction. It is thought that these subspecies also correspond to multiple waves of ancestral wolves that invaded North America via the Bering land bridge, after their likely predecessor, *Canis mosbachensis* evolved into *Canis lupus* in the Old World (Nowak 2003). This re-colonization pattern has resulted in "centrifugal evolution," (*sensu* Groves 1993) with subspecies showing increasing morphological divergence the farther they are from the point of entrance at Alaska (Nowak 2003). Further, Nowak (2003) suggests that some New World subspecies are more closely related to some of the Old World subspecies than to certain subspecies on their own continent, and vice versa. While grey wolves evolved into their current form in Eurasia, *Canis lepophagus*, which is thought to be the most recent predecessor of modern wolves and coyotes (Kurtén and Anderson 1980), evolved originally in North America before some individuals travelled to Eurasia.

Grey wolf characteristics and life history

In North America, grey wolves inhabit Canada, Alaska, and a few small areas in the contiguous USA, with Canada containing the greatest number of wolves of any country, an estimated 50-60 000 (Boitani 2003). Wolves are the largest of all canids, and can weigh as much as 62kg (Mech and Boitani 2004b). Wolves display diverse pelage colour patterns, with a continuous spectrum ranging from black to white, which may also include brown, grey, and red hairs (Gipson *et al.* 2002; Mech and Boitani 2004b). Coat colours have been observed to vary significantly with geography, and a distinct northeast/southwest cline has been observed in several studies, following an environmental gradient of arctic tundra to boreal forest in northern Canada (Gipson *et al.* 2002; Musiani *et al.* 2007). The tundra contains a significantly higher percentage of white and light coloured wolves than the forest, which shows increasing numbers of black or dark

individuals to the south and west. This colour differentiation is believed to have a selective advantage, better allowing wolves to blend in with their local surroundings and thus making them harder to detect by prey (Jolicoeur 1959).

Grey wolves exhibit a variety of diets, and will catch prey ranging in size from snowshoe hares (*Lepus americanus*) to bison (*Bison bison*), depending on temporal and geographic availability. While stable wolf populations may require the presence of large ungulates for sustenance (Mech 2005), and most of their diet consists of prey, wolves are not obligate hypercarnivores. They have even been seen scavenging through garbage (Peterson and Ciucci 2003), although such behaviour is more typical of coyotes, which, unlike wolves, tend to thrive near human settlements (Mech and Boitani 2004a).

Wolves are a highly social species, and are often found in packs that contain as many as 42 individuals (Mech and Boitani 2003). Packs usually consist of a dominant breeding pair of individuals, their offspring, and other relatives that either originated in other packs or remained with their natal pack into adulthood, although almost all wolves will eventually disperse from their natal packs (Mech and Boitani 2003). Packs are usually very territorial, and can occupy home ranges as large as 4300 km^2 (Mech *et al.* 1998), depending on pack size and availability of food (Mech and Boitani 2003). However, some wolves that feed primarily on migratory prey such as barren-ground caribou (*Rangifer tarandus groenlandicus*; e.g. Parker 1973; Walton *et al.* 2001) or saiga antelope (*Saiga tatarica*; Mech and Boitani 2003) migrate with their prey and do not maintain permanent territories. While seasonal migration events can occur over > 500 km (Walton *et al.* 2001), individual wolves have been observed to travel > 800 km when dispersing away from their natal packs (Fritts 1983; Wabakken *et al.* 2007), although dispersal distances are typically significantly less than this and likely depend on the availability of unoccupied territory (Mech and Boitani 2003).

Despite their ability to disperse such large distances, grey wolves typically exhibit significant population genetic structure at continental scales, even when sampled across a continuous distribution (see Pilot *et al.* 2006; Carmichael *et al.*

2007). They are significantly more structured than the only other circumpolar canid, the arctic fox (*Alopex lagopus*; Carmichael *et al.* 2007). Such population structure, however, does not appear to be a result of simple geographic distance, since neither study found evidence of strong of isolation by distance (IBD) in grey wolves even at very large scales (Pilot *et al.* 2006; Carmichael *et al.* 2007). Rather, differing ecology seems to predict population structure significantly better than simple geographic distance. Pilot *et al.* (2006) found that population structure was strongly correlated with climate, habitat, and diet in eastern European wolves. Similarly, Carmichael *et al.* (2007) and Geffen *et al.* (2004) found that habitat type explained a significant amount of variation in genetic distance between subpopulations of wolves in North America, and Musiani *et al.* (2007) found that prey type (migratory vs. non-migratory) and habitat type (boreal forest vs. arctic tundra) explained a significant proportion of genetic variation at a smaller scale within North America.

SNP identification and genotyping

For the past fifteen years, population geneticists have relied primarily on microsatellite markers in order to assess population structure, genetic variability, census sizes, and paternity, as well as to map quantitative trait loci. Microsatellites have many features to recommend them for such uses: they are relatively cheap and quick to develop and analyze, they are abundant in mammals and fish (Neff and Gross 2001), they often amplify across species, they can be highly polymorphic and they are generally considered selectively neutral (Vignal *et al.* 2002; although this assumption is often incorrect). Additionally, there are several available models of evolution designed specifically to reflect microsatellite mutation, including the stepwise mutation model (Kimura and Ohta 1978) and the two-phase mutation model (di Rienzo *et al.* 1994). However, alongside all the benefits of using microsatellites, there are a number of drawbacks. First, when evaluating microsatellites, it is not possible to account for unexpected mutations. For instance, because of their high mutation rate, there is a risk of size homoplasy

(Pompanon *et al.* 2005): since microsatellite alleles are evaluated based on size, it is assumed that alleles of the same length share identical sequence and are identical by descent. Both of these assumptions can be broken without detection, leading to erroneous conclusions. Alternatively, if there is a mutation near the 3' end of the priming site of one allele, this allele will not readily amplify, and individuals with such alleles will appear to be homozygous for their alternate allele. Further, there is the issue of allelic dropout, whereby shorter alleles are preferentially amplified, which can make heterozygous individuals appear to be homozygous at the shorter allele. Additionally, due to the high allelic polymorphism of most loci, microsatellites are prone to sampling error in that they may provide an inaccurate estimate of allele frequencies in data sets with a limited number of samples (Ruzzante 1998), yielding results that will complicate accurate assessment of population structure. There is also the issue of ascertainment bias: since loci with high polymorphism are selected for use, rather than randomly selected loci, heterozygosity at microsatellite loci is not a good predictor of overall genetic diversity, as assessed by re-sequencing introns (Väli *et al.* 2008).

Because of the many drawbacks of using microsatellites, the decreasing cost of large-scale genetic analyses, and the increasing availability of genomic resources, it is desirable to find a marker that is able to overcome some of the limitations of microsatellites, and single nucleotide polymorphisms (SNPs) have been suggested as a marker that should prove to be highly useful in population genetics (Morin *et al.* 2004). In brief, a SNP is a single variable nucleotide site within a species' genome, usually having only two alleles. Ideally, the less common allele should occur at a frequency that is $\geq 1\%$ (Vignal *et al.* 2002), indicating that the polymorphism persists within the species and is not simply a chance mutation that will be subsequently lost. SNPs have several advantages over microsatellites for use in genetic studies. First, SNPs are probably the most common genetic polymorphism (Brumfield *et al.* 2003), and therefore the most abundant marker in the genome, allowing for a greater density of markers for use in genome-wide analyses such as population genomics or disease association

studies, and more accurately representing genetic diversity within a genome. Further, SNPs can be processed at a much higher throughput level than microsatellites. Using microarray technology, ~500 000 SNPs can be genotyped in an individual at once (e.g. Novembre *et al.* 2008), ~4 orders of magnitude greater than is possible for simultaneous microsatellite genotyping. Second, the mutational process surrounding SNPs is conceptually simpler. While there are several models appropriate to explain microsatellite evolution, it is difficult to assess which model is most appropriate for a given data set. There is only one model commonly used to explain SNP evolution, the infinite site model proposed by Kimura (1969), whereby a mutation occurs only once at any nucleotide site. This model is highly appropriate because SNPs have a low mutation rate ($\sim 10^{-8}$ per generation), so usually have only two possible states and are unlikely to undergo mutational reversals (Brumfield *et al.* 2003). Third, because of their low allelic polymorphism, SNPs do not suffer from sampling error (see above), which should make them better for detecting population structure in a limited number of individuals. However, this means that each locus is less informative, and in fact simulations suggest that up to 10x more binomial markers than multi-allelic markers (such as microsatellites) may be required to accurately estimate genome-wide variability (Mariette *et al.* 2002). That said, this lack of power is easily compensated for through the far greater number of SNPs that can be genotyped. Additionally, because the genomic coverage offered by SNPs is so great, they are an ideal marker to use when looking for statistical signatures of selection, or markers associated with a specific phenotype.

Before SNP profiles can be evaluated in any organism, multiple individuals must be (partially) sequenced in order to locate SNPs in the genome, and to determine flanking sequence so they can be individually interrogated. There are numerous methods for accomplishing this, all of which involve either targeted sequencing or random sequencing of genomic regions (Slate *et al.* 2009). For studies of non-model species, where there is little available sequence data, and likely limited funding, the exon-priming intron-crossing (EPIC) targeted sequencing approach is commonly used (e.g. Aitken *et al.* 2004). This is done by

designing primers from the ends of neighbouring exons (determined from sequence of a closely related species, and thus likely conserved), and using them to amplify and sequence the intronic regions between them in the target species. Introns are more likely to contain SNPs than exons, because they are generally less functionally constrained (Slate *et al.* 2009), but of course whether coding SNPs are preferred or not will depend on what questions they will be used to address. Another increasingly common method for SNP detection involves the random sequencing of pooled cDNA (DNA derived from reverse transcription of mRNA) from multiple individuals on a high-throughput sequencer, such as the 454 GS FLX, in order to generate expressed sequence tags (ESTs). Due to the high degree of coverage obtainable with the 454 sequencer, SNPs within protein-coding genes can easily be detected by comparing overlapping sequences using specialized software such as PolyBayes (Marth *et al.* 1999) or QualitySNP (Tang *et al.* 2006). Because hundreds or thousands of genes are sequenced simultaneously, this approach can yield many more SNPs than the EPIC approach, up to many thousands of SNPs at a time (e.g. Gore *et al.* 2009). Alternatively, if no high-throughput sequencer is available, sequencing of randomly generated amplified fragment length polymorphisms (AFLPs) can be done in order to discover SNPs in any organism, regardless of the availability of sequences or other genetic tools (e.g. Roden *et al.* 2009). Ultimately, the method used for discovering SNPs will be based on three main factors: 1) availability of previously generated sequence data, 2) number, location (genic vs. non-genic), and type (coding vs. non-coding) of SNPs desired for research, and 3) available funds and tools.

Whichever method is used for discovering SNPs, one needs to be aware of a potential problem when locating SNPs: ascertainment bias, which refers to a bias in how SNPs are detected. For instance, if a group of SNPs to be used to assess genetic variability is screened from one population, researchers may miss SNPs that are fixed in this population, or are at too low a frequency to detect. Such a bias will alter the frequency spectrum, the distribution of frequencies at which SNP alleles occur in a group of individuals (Nielsen *et al.* 2004). This can

affect results obtained when trying to assess overall genetic diversity, divergence times, migration rates, population structure, or when looking for signals of selection (Morin *et al.* 2004; Nielsen *et al.* 2004). However, ascertainment bias can be avoided by using samples from all populations that will be studied when generating SNPs, and it is sometimes possible to correct for ascertainment bias, if enough is known about the ascertainment scheme (Brumfield *et al.* 2003).

Once it has been decided which SNPs will be interrogated for a given project, a method for genotyping must be decided upon. Like SNP discovery, many possible methods for SNP genotyping exist. If relatively few loci are to be genotyped, there are several methods for genotyping a single SNP at a time. A good example is the TaqMan assay (from Applied Biosystems), performed with a real-time PCR thermocycler. Because the TaqMan assay is PCR-based, it can be used for multiplexing in order to genotype several (<10) SNPs in one reaction. At the opposite end of the spectrum is genotyping using microarrays that can be purchased, either pre-made or custom-designed, from companies such as Affymetrix and Illumina. As previously mentioned, microarrays can genotype many thousands of SNPs at once. Each microarray is divided into thousands of partitions, with each partition containing the probe for one SNP allele. By adding a fluorescent dye to amplified DNA, hybridized DNA will give off light where it hybridizes, and by comparing the amount of light given off at each site on the array, individuals can be genotyped concurrently at all SNP loci being interrogated. Unfortunately, because so many SNPs are genotyped at once, hybridization conditions are not specific to each SNP, so some SNPs hybridize more easily than others. In order to minimize the mis-calling of genotypes due to nonspecific hybridization, there are usually multiple copies of each probe, spread out over different locations on the array, to try to reduce any bias in hybridization. While the cost per SNP for genotyping is much lower for microarrays than with most other SNP genotyping methods, microarrays are nonetheless very expensive, and require specialized equipment in order to hybridize DNA to them, and to scan afterwards in order to evaluate fluorescence. As is the case with SNP detection, the appropriate method for SNP genotyping will vary with each project,

depending on the number of SNPs to be genotyped, as well as available time and funds.

Population genomics

Traditional population genetics is the study of microevolutionary forces, namely mutation, genetic drift, selection, and gene flow, assessed by comparing allele frequencies of neutral markers within species and populations. This has been done for many years using a relatively small numbers of loci that require little or no development, such as restriction fragment length polymorphisms (RFLPs), later followed by amplified fragment length polymorphisms (AFLPs), and eventually microsatellite loci, which require additional time and money to develop, but are multi-allelic and co-dominant, and so offer greater discrimination on a per marker basis. Population genomics is an extension of this kind of analysis, using a large amount of genome-wide genetic data (such as SNPs, or possibly AFLPs) to answer population genetics questions more specifically and robustly. An important distinction between the two fields is that with the large amount of data used, population genomic studies have the power to distinguish genome-wide genetic effects (such as bottlenecks and inbreeding) from locus-specific effects (such as mutation, recombination, and selection) (Black *et al.* 2001; Luikart *et al.* 2003; Stinchcombe and Hoekstra 2007). This is highly important because only neutrally evolving loci (i.e. loci that exhibit genome-wide patterns) will accurately inform estimates of demographics, population structure and migration, and having a set of neutral loci allows for a comparative assessment of mutation and recombination, as well as statistical detection of selection in loci that do not appear to be evolving neutrally.

Luikart *et al.* (2003) suggested a thorough schematic for conducting a population genomics study, which can be simplified to the following general methodology: sample many individuals and genotype them at many loci, look for outlier loci using statistical tests, estimate demographic parameters using the

neutral loci, and determine causality of outlier loci. The main issues to note with sampling are that, depending on the study questions, a large number of samples is preferable and often requisite, and individuals should be sampled at regular geographic intervals, without using prior assumptions about populations or breeding groups. Genotyping should be done with as many loci as possible, spaced broadly throughout the genome. Isolating outlier loci from the rest of the data set is often an analytically challenging aspect of a population genomics study, so I will briefly discuss how outlier loci, or loci that are candidates for selection, are identified.

An increasingly used approach to take when looking for functional genotypes is to look for loci that exhibit statistical evidence of selection, rather than looking for associations between specific phenotypes and genotypes. There have been several methods proposed for doing this, the simplest of which is to look for loci that are outliers in between-group F_{ST} distributions. The theory, postulated by Lewontin and Krakauer (1973), is that loci linked with genes undergoing directional selection should differ highly in allele frequencies between groups in comparison to neutrally evolving loci, or loci linked to genes undergoing balancing selection. Of course, there is no guarantee that outliers are under directional selection, as high F_{ST} values may simply be caused by genetic drift between groups, but this method will highlight some loci that merit further investigation. For studies with few genetic markers, there is available software that simulates a null distribution of F_{ST} values, such as Beaumont and Nichols's (1996) FDIST. However, this is not ideal, because the simulated null distribution can be biased by poorly estimated population demographics or structuring (Excoffier *et al.* 2009), and so it is preferable to use a large set of markers in order to obtain an empirical distribution of F_{ST} values. This allows for an empirical (*sensu* Teshima *et al.* 2006) genome scan.

There are two other main types of methods to look for selection in a set of loci, based on the frequency spectrum and linkage disequilibrium (LD) respectively. Tajima (1989) first proposed the use of the frequency spectrum to

identify loci that deviate from neutrality, measured with a statistic that he proposed (*d*). He hypothesized that loci that are in LD with a selected gene or marker would exhibit low minor allele frequencies relative to other loci in the genome. The selected locus would drag individual alleles at nearby neutral loci into high frequency, reducing the frequency of the other alleles at these loci. Such low-frequency neutral alleles would return to higher frequency only after recombination broke up that particular linkage block over many generations. However, using Tajima's (*d*) to look for selection is unadvisable if the demographic history of the study organism is unknown, because it is difficult to distinguish between recent selection and a recent bottleneck using this test: both scenarios have the effect of reducing the number of available alleles in past generations (Simonsen *et al.* 1995). Another caveat is that the frequency spectrum cannot be used to detect selection in loci that have swept to fixation, because a monomorphic locus is not necessarily indicative of directional selection. Finally, Tajima's (1989) *d* will have little power to detect selected loci in systems where there is strong ascertainment bias, because as previously mentioned, ascertainment bias can result in relatively few loci with low minor allele frequencies being used in the study system, which will skew the results.

Fortunately, the LD-based methods are not affected by a skew in the frequency spectrum, although they are still sensitive to assumptions about demographics. LD methods, such as the integrated haplotype score (iHS; Voight *et al.* 2006) or the linkage disequilibrium decay (LDD) test (Wang *et al.* 2006) look for areas in the genome that exhibit high LD, under the assumption that these areas are undergoing positive selection which would cause increased LD. Because LD tends to decay over time, LD-based analyses work best at finding loci that have undergone recent selection, although they can still detect selection at recently fixed loci. A potential problem is that most LD-based analyses require knowledge of the phase of the genotypes, that is, knowing which chromosome from a chromosomal pair each allele is on. While this is not an issue when using sequence data, phase is not known for genotype data. However, phasing can be estimated using maximum likelihood (Excoffier and Slatkin 1995) or Bayesian

(Scheet and Stephens 2006) frameworks. Lastly, most LD-based methods rely on an estimate of the underlying recombination rate (Nielsen 2005), and if this rate is inaccurate, then the results from these tests may be unreliable.

No matter which method is used to detect loci under selection, the results obtained usually require further investigation on an individual level. For those lucky enough to have a sequenced genome available for a closely related species, the genome browser provided by the Genome Bioinformatics group of UC Santa Cruz (http://genome.ucsc.edu/cgi-bin/hgGateway) can be valuable for examining the genomic context of markers that show evidence of selection. For instance, if you know the location of a SNP, you can search for it in the genome browser and look for nearby genes, which may give some indication of what phenotype the detected locus could be affecting. Alternatively, if such resources are unavailable, you could BLAST (Altschul *et al.* 1990) the sequence surrounding the marker in question (if known) at NCBI, to see if it has any significance in another study system.

Thesis goals

The theoretical benefits of using SNP loci rather than microsatellites for population genetic studies seem clear: much higher genomic coverage, the ability to more accurately detect population structure from small sample sizes, and the ability to discover loci that have undergone, or are undergoing, selection. However, it remains to be empirically documented what the benefits of using a large SNP data sets are when studying wild populations. To date, most large-scale SNP studies (i.e. studies with > 10 000 loci) have involved human populations, where there are hundreds of available samples, and 500K+ markers, ascertained within and between human populations. For non-humans, large-scale SNP studies have focussed on economically important organisms such as cattle (*Bos primigenius*; e.g. Pant *et al.* 2010), domestic dogs (vonHoldt *et al.* 2010) and domestic sheep (*Ovis aries*; Kijas *et al.* 2009), which also have many available

samples, and many thousands of SNPs ascertained within each species. While studies of wild populations have also been conducted using SNP arrays, no within-species comparison has approached the scale of the studies of domesticated animals. Generally, wild animals have been genotyped on arrays developed for a related domestic species, so the number of useable, polymorphic SNPs is dramatically reduced [e.g. 2-3% of SNPs in bison genotyped on a cattle SNP array, (Pertoldi *et al.* 2009) and wild sheep species genotyped on a domestic sheep array, (Kijas *et al.* 2009)]. Thus, it has not yet been possible to evaluate the benefit of possessing a large SNP data set for determining population structure, since no wild population has been genotyped at a large number of markers, and all domestic animals have been intentionally segregated. Additionally, because studies of wild populations generally have fewer loci, most studies looking for selection have had to rely on simulated null distributions in order to detect outliers.

I aim to expand on what has been learned about wolf population structure from microsatellite studies by genotyping a wild population of wolves using the Affymetrix Canine Mapping Array version 2. This array contains probes for ~127 000 SNP loci, and has already been used to successfully genotype a global panel of > 200 grey wolves and > 900 domestic dogs at ~48 000 SNP loci, in order to infer the primary location of dog domestication (vonHoldt *et al.* 2010). Because wolves and dogs are so closely related (see above), a much larger proportion of loci can be successfully genotyped on this array in grey wolves than has been the case for other cross-species SNP array studies.

Using this array, I genotyped grey wolf samples from across North America, in order to re-assess population structure using a high-density marker. This will allow me to compare the results obtained from large-scale SNP typing to those obtained from the panel of 14 microsatellites used by Carmichael *et al.* (2007). Using this approach in Chapter 2, I pursue three goals: 1) expand understanding of grey wolf population structure through the use of a high-density genetic marker, 2) outline the differences observed through the use of many SNP

markers rather than a small microsatellite data set and 3) evaluate how many SNPs are required in order to accurately assess population structure.

In Chapter 3, I look for evidence of positive selection in this population, in the hope of identifying genes that are important to phenotypic variation within North American grey wolves. Because wolves live in diverse habitats (Geffen *et al.* 2004), and population structure seems to be highly influenced by ecological factors (see above), it is likely that wolves in different subpopulations possess habitat-specific genetic adaptations to their surroundings. Because I am working with a large set of SNP markers, I perform an empirical genome scan to look for SNPs that are under selection between subpopulations. I look at the extreme outliers of the distribution and identify SNPs that may be under selection and identify nearby genes. Finally, I discuss why these genes might be important for phenotypic differentiation across this range.

Chapter 2

Population structure of North American grey wolves

Introduction

With the development of molecular markers, the field of empirical population genetics has taken off in the past 30 years. More recently, the ease of development, high polymorphism, and low cost of microsatellite markers has enabled the study of many non-model species, which in turn has yielded much insight into microevolutionary processes. Beyond illuminating dispersal, mating systems and large-scale phylogeographic patterns, population genetics can also be applied to epidemiology for monitoring the development and spread of pathogens (e.g. Nübel *et al.* 2010), and for livestock breeding programs through the mapping of quantitative trait loci.

Large-scale studies of human populations are becoming increasingly common, with researchers genotyping hundreds of individuals at hundreds of thousands of SNP loci. The amount of data collected dwarfs what was possible only a few years ago (e.g. ~197 000 SNPs; Novembre *et al.* 2008), and progressively larger panels of markers are still being created (e.g. the Affymetrix Genome-wide Human SNP Array 6.0, with > 900 000 SNPs and > 900 000 probes for copy number variation). It was the sequencing of the human genome (Venter *et al.* 2001) that has allowed for the development of these vast sets of markers, and such large numbers of markers are not available for any other species. However, sequencing efforts continue in model and non-model species, so the genomic resources for other species are ever increasing.

The sequencing of the domestic dog genome (Lindblad-Toh *et al.* 2005) enabled the development of a canine SNP microarray containing ~ 127 000 SNPs, which has recently been used to examine the genetic diversity and infer origins of domestic dogs (vonHoldt *et al.* 2010), and this same array can also be used to

successfully genotype species as divergent as coyotes, jackals, and African wild dogs (vonHoldt *et al.* in prep). But while these studies aimed to quantify species differences and evolutionary history, they were not designed to evaluate within-species genetic variation in relation to natural population distribution.

In this chapter I made a preliminary assessment of the benefits of using a large marker set instead of a small set of microsatellites for studying population structure in grey wolves. I have genotyped > 130 wolves sampled from across northern Canada and Alaska using these canine SNP microarrays, and I used this data to re-evaluate the population structure of North American wolves. I made basic comparisons between results from the SNP data and previous results from microsatellite data, looking at isolation by distance, population structure and differentiation within and between subpopulations, exploring biological and theoretical reasons behind discrepancies. I also made comparisons between the observed population structure and previously proposed subspecies distributions.

Additionally, I evaluated the relative performance of different numbers of SNP markers when assigning individuals to subpopulations. STRUCTURE (Pritchard *et al.* 2000) is a commonly used computer program that assigns individuals to genetic clusters and evaluates admixture between these clusters, but there is currently no objective method with which to evaluate the resulting individual assignments. By comparing the results from STRUCTURE runs using different-sized subsets of SNPs, I provided an estimate of the number of SNP markers needed to assign individuals accurately and to precisely evaluate admixture.

Methods

Sample selection and genotyping

The samples that were genotyped were selected from a set of > 2000 grey wolves used in previous studies (Carmichael *et al.* 2007; Musiani *et al.* 2007) with

an additional 30 tissue samples obtained from the University of Alaska Museum (Fairbanks, AK), based primarily on location, and secondarily on the quality of tissue. 45 samples were previously genotyped on a customized genome-wide 127k Affymetrix Canine Mapping Array (version 2) for use in a study of dog domestication (vonHoldt *et al.* 2010). To supplement these samples, an additional 110 wolves were selected as candidates for genotyping with the intention of maximizing the breadth of sampling across northern Canada and Alaska. In addition to the above criteria, ten wolves from the Alexander Archipelago in the Alaskan Panhandle were selected for genotyping because wolves on these islands have been previously defined as a subpopulation distinct from mainland wolves in adjacent mainland British Columbia (Weckworth *et al.* 2005; Muñoz-Fuentes *et al.* 2009). Ten wolves were selected so that wolves from this area were adequately represented in the data set.

DNA was extracted using a QIAamp DNA mini kit (QIAGEN) following standard protocol and quantified using a Nanodrop 1000 (Thermo Scientific, Wilmington, Denver). Samples with 260/280 readings < 1.8 or an insufficient quantity of DNA were discarded, and samples with 260/230 readings < 2.0 were cleaned up by ethanol precipitation prior to processing. 87 samples (plus 5 duplicates) were genotyped on the SNP array following the Affymetrix "GeneChip® Mapping 500K Assay" protocol. Before the hybridization step, sample volumes were reduced to 35μL by heated evaporation in order to allow the entire volume of each sample to be hybridized to a single array. Samples that could not be successfully prepared (due to low concentration of DNA or failure during preparatory procedure; n = 23) were discarded prior to array hybridization.

After hybridization and scanning, genotypes were called using BRLMM-P software provided by Affymetrix (Affymetrix technical report; http://www.affymetrix.com/support/technical/whitepapers/brlmmp_whitepaper.pdf). In order to ensure only reliable genotype calls were used in the analysis, three quality control measures were taken. SNPs with < 95% call rate, heterozygosity > 70%, or minor allele frequency (MAF) < 5% across the entire data set were

removed. All X-chromosome SNPs (n = 1996) were discarded because the data set included males. A total data set of 27 931 SNPs were retained for analysis. To evaluate consistency of protocol and genotyping calls, five samples were fully processed and genotyped in duplicate. Genotype calls between duplicated samples were found to differ at < 1.2% of SNP loci, with > 80% of the discrepancies involving an uncalled genotype in one sample rather than a disagreement between called genotypes. Because of the potential for linkage disequilibrium to bias results, sliding windows of ten SNPs within each chromosome were evaluated for high correlation ($r^2 \geq 0.5$) using PLINK v.1.06 (Purcell *et al.* 2007). If any pair of SNPs in any 10-SNP window was observed to have $r^2 \geq 0.5$, one SNP was randomly removed by PLINK. This pruning yielded a reduced data set of 26 221 SNPs (henceforth referred to as LD-pruned) that are not in high LD due to close proximity.

Population structure

In order to determine the population structure of North American grey wolves, STRUCTURE (Pritchard *et al.* 2000; Falush *et al.* 2003) was used to identify genetic clusters of individuals. STRUCTURE uses a Bayesian Markov chain Monte Carlo (MCMC) algorithm to group individuals into a pre-determined number of clusters (K), and the likelihood of the result [Ln P (D|K), where D is the genotype data] is evaluated, and can be compared between runs using different values of K. Under the admixture model, admixture proportions are then obtained for each individual from the run with the highest Ln P (D|K). For this analysis STRUCTURE was run with 20 000 burn-in iterations and 5 000 sampling iterations of the MCMC for K = 1 through 10 with correlated allele frequencies (as recommended by the authors) for the entire set of samples using the LD-pruned data set. The admixture ancestry model was selected, which allows different genes within an individual to be inherited from different genetic clusters. This is the preferred model for this study system because of grey wolves' known ability to migrate long distances and because the sampling of individuals across

the landscape was evenly distributed, with few obvious *a priori* subpopulation boundaries or barriers to gene flow. Each run was performed five times, and the run with the highest Ln P (D) for each *K* value was compared in order to determine the best value of *K*, as suggested by Pritchard *et al.* (2000) and Faubet *et al.* (2007). STRUCTURE was then re-run three times with *K* fixed at this value with the number of sampling iterations increased to 10 000 in order to recover the best admixture proportions, and the run with the highest Ln P (D) was selected. Individuals that had > 75% assignment to a single genetic cluster were considered to be part of a subpopulation.

In addition to the above Bayesian method examining population structure, I performed a principal components analysis (PCA; Patterson *et al.* 2006). PCA provides a quick way of summarizing multi-dimensional data into more easily interpreted components, which can then be plotted on a scatter chart. This technique has been used extensively in recent human genetics literature (e.g. Lao *et al.* 2008; Novembre *et al.* 2008; Bishop *et al.* 2009; Silva-Zolezzi *et al.* 2009), where studies include hundreds or thousands of individuals genotyped at 500K+ SNP loci, because it is much quicker than computation-heavy Bayesian clustering algorithms like STRUCTURE. To perform PCA, I used the program SMARTPCA within the EIGENSTRAT software package (Price *et al.* 2006) using the 132-individual, LD-pruned SNP data set. The values for the first two principal components were plotted to identify distinct groups of individuals. Subpopulations determined using STRUCTURE were superimposed to see if they were confirmed by PCA.

To measure the degree of genetic differentiation between clusters I calculated Weir and Cockerham's (1984) θ, analogous to Sewall Wright's (1951) F_{ST} (and henceforth referred to as F_{ST}), across all clusters and between each pair of clusters using a script written by J. Novembre. To assess the amount of variation in inter-individual identity-by-state genetic distances (D_{IBS}; calculated using PLINK) within subpopulations vs. between subpopulations, I performed an analysis of molecular variance (AMOVA) using ARLEQUIN v.3.11 (Excoffier *et*

al. 2005). A total of 10 000 permutations of the data set were performed to assess significance.

Subpopulation structure

While the number of individuals used in this study was small relative to the scale of the sampled area, because so many markers were used I wished to determine whether it was possible to detect finer population structure within subpopulations. To investigate this, I ran STRUCTURE analyses on each subpopulation determined from the previous step. Data subsets consisting of all individuals assigned to each subpopulation were input into separate STRUCTURE runs with *K* values of 1-5, to see if further structure could be detected. As before, the LD-pruned data set was used, and STRUCTURE was run with the admixture model and correlated allele frequencies. Because sample sizes were reduced compared to running the entire population at once, STRUCTURE was run five times for 20 000 burn-in cycles with 20 000 subsequent iterations, with the run scoring the highest Ln P (D|*K*) used to select the most appropriate *K* value.

Isolation by distance

To look for evidence of isolation by distance (IBD), Mantel tests (Mantel 1967) were performed to compare D_{IBS} with pair-wise geographic distances calculated using GENALEX v.6 (Peakall and Smouse 2006). The Mantel analysis was performed with the *vegan* v. 1.15-4 package (Oksanen *et al.* 2009) in R (R Development Core Team, 2009) using 1000 permutations to test the correlation between genetic and Log_{10}–transformed geographic distance.

To make comparisons with microsatellite data, a subset of wolves also genotyped with 14 microsatellite loci by Carmichael *et al.* (2007) was analyzed separately with a Mantel test. Then, an allele-sharing distance (analogous to D_{IBS})

between each pair of these 61 individuals was calculated from the microsatellite data using the MICROSATELLITE TOOLKIT (Park 2001). A separate Mantel test was run using the allele-sharing distances for these 61 individuals.

To evaluate the relative effect of geographic distance compared to population structure on D_{IBS}, partial Mantel analyses (Smouse *et al.* 1986) were performed. For these analyses I used only individuals assigned to a single subpopulation. I constructed an additional distance matrix, where individuals assigned to the same subpopulation were coded to have a distance of 0, and those from different subpopulations were coded with a distance of 1. Two sets of analyses were run. In the first, subpopulation assignment distance was held constant, in order to evaluate the importance of geographic distance to D_{IBS} after accounting for population structure. In the second, geographic distance was held constant, in order to determine the importance of population structure on D_{IBS} after accounting for geographic distance.

Additional Mantel analyses were run for all individuals within each subpopulation to look for evidence of IBD on a smaller scale and within similar environments. For clusters containing wolves separated by large water barriers (segments of ocean, but not rivers), partial Mantel analyses were also performed. To do this, additional distance matrices were constructed to control for whether each pair of wolves was separated by a water barrier (distance = 1) or not (distance = 0), so that the effect of a water barrier would not influence the correlation between geographic and genetic distances.

Comparison of SNP data sets and microsatellites for determining population structure

To determine the relative ability of different amounts of SNPs to correctly assess population structure, ten random subsets of each of 5000, 1000, 500, 140, 98, and 56 SNPs were generated. The latter three numbers were picked because they are ten, seven, and four times (respectively) the number of markers used by

Carmichael *et al.* (2007) to assess population structure, and so could be used for comparison purposes with results from microsatellite data. All 132 wolves were run in triplicate in STRUCTURE with only the genotype data from each SNP subset and with a fixed *K* value (determined during assessment of population structure; see above). This allowed comparison between the results obtained from SNP subsets to those from the complete LD-pruned data set, under the assumption that the latter data set would provide the most accurate results. Due to the reduced number of markers, each size of data subset was run for an increased number of burn-in iterations (50 000, 100 000, 200 000, 400 000, 500 000 and 1 000 000 iterations respectively) and sampling iterations (50 000, 200 000, 500 000, 1 000 000, 1 200 000, and 2 000 000 iterations respectively). Amongst each triplicate set of runs, assignments for each subset were taken from the replicate with the highest Ln P (D). These assignment matrices were then run alongside the LD-pruned 26K assignment matrix in the program CLUMPP (Jakobsson and Rosenberg 2007) in order to permute the cluster assignments (which are randomly labelled in STRUCTURE) to match across all the runs. For this calculation, the GREEDY algorithm was selected (to minimize time consumption vs. running a complete search) with 10 000 random input orders. Finally, two metrics were used in order to determine the ability of each subset of SNPs to correctly identify population structure. The first measure I calculated was the "clusteredness" (*G*) of each individual within each run, as defined by Rosenberg *et al.* (2005). Clusteredness measures the degree to which individuals are assigned to a single cluster (clusteredness = 1) rather than being assigned equally to all possible clusters (clusteredness = 0) and is calculated across a group of individuals as

$$G = \frac{1}{I}\sum_{i=1}^{I}\sqrt{\frac{K}{K-1}\sum_{k=1}^{K}(q_{ik} - 1/K)^2}$$

where I is the total number of individuals, K is the total number of clusters, and q_{ik} denotes the admixture coefficient of the ith individual in the kth cluster. Compared to just using the highest assignment proportion of each individual, the clusteredness value is reduced if the remaining assignment is evenly broken up amongst other clusters. This distinction is significant because when STRUCTURE is unable to detect population structure in a data set (either due to lack of power or a true lack of structure), it will tend to assign each individual equally to all K clusters (Rosenberg *et al.* 2005). Additionally, this measure corrects for K so that K does not directly influence clusteredness values. Clusteredness was calculated for each individual in each individual SNP subset, and then averaged across all subsets with the same number of markers.

A second metric was devised to evaluate the similarity between the results from any data subset and the complete 26k LD-pruned data set. It was calculated similarly to the clusteredness, with minor modifications, whereby the admixture coefficient from the complete LD-pruned data set was used instead of the $1/K$ term, and the $\frac{K}{K-1}$ term was removed. It will be referred to as discordance (d), and was calculated across individuals as

$$d = \frac{1}{I}\sum_{i=1}^{I}\frac{\sqrt{\sum_{k=1}^{K}(q_{iks} - q_{ikc})^2}}{\sqrt{2}}$$

where q_{iks} is the admixture coefficient of the ith individual in the kth cluster estimated using a data subset (s), and q_{ikc} is the admixture coefficient of the ith individual in the kth cluster using the complete LD-pruned 26k data set (c). The main term is divided by $\sqrt{2}$ in order to scale the measure between 0 and 1. Discordance is calculated in such a way that mis-assignment is more highly

penalized if it is concentrated in a single cluster, rather than broken up amongst several clusters. Thus, individuals that are strongly assigned to a to a single incorrect cluster show greater discordance than individuals that incorrectly appear to have a mixed genetic background.

Next, the subset of wolves also genotyped by Carmichael *et al.* (2007) (n = 61; henceforth referred to as the 61-individual data set) was run in STRUCTURE (with admixture and correlated allele frequencies) with the LD-pruned SNP set for $K = 1$ to 7. As before, five replicates were run for each value of K in order to determine the most appropriate value. Then, this subset of individuals was run in triplicate in STRUCTURE with K fixed for each of the SNP subsets, using the same number of burn-in and sampling iterations as with the 132-individual data set. Last, the microsatellite data for the 61-individual data set was run 3 times in STRUCTURE with K fixed for 1 000 000 burn-in iterations and 2 000 000 sampling iterations, in order to evaluate how well this microsatellite data set was able to determine population structure relative to SNP data sets of varying sizes. Discordance and clusteredness were calculated for the runs with the highest Ln P (D) for each SNP data subset and for the microsatellite data.

Finally, I tested the correlation between clusteredness and discordance, by comparing the mean discordance for each individual across subsets of the same number of SNPs, and the clusteredness score of each individual assessed using the complete LD-pruned SNP set (henceforth referred to as best-estimate cluteredness). To evaluate the significance of the correlation, Spearman's rank correlation coefficient was calculated for each comparison. This was also done for the microsatellite results.

Results

Population structure

STRUCTURE runs of the entire set of individuals reached an asymptote in Ln P (D|*K*) at $K = 5$ (Fig. 2-1). I therefore used $K = 5$ genetic clusters for subsequent analysis. The five clusters are geographically coherent (see Fig. 2-2a), and increasing *K* past 5 yielded no additional clusters to which individuals assigned strongly (i.e. ≥ 75%). Five subpopulations (totalling 87 wolves) containing individuals that assigned ≥ 75% to a single genetic cluster were named according to their geographic origin: Pacific, Forest, Arctic, Baffin Island, and Atlantic. All five subpopulations were geographically discrete (Fig. 2-2b).

Using PCA, the same geographically coherent groups appeared distinct according to their scores on the first two axes, PC 1 and PC 2 (Fig. 2-3). The first and second axes accounted for 5.6% and 4.3% (respectively) of the observed genetic variation. There was high congruence between the STRUCTURE subpopulation assignments and their pattern of clustering by PCA. Unassigned individuals were generally intermediate between the subpopulation pairs most highly represented in their genomes.

Genetic differentiation measured among all subpopulations was moderate with global $F_{ST} = 0.131$. Pair-wise F_{ST} ranged from 0.065 between Forest and Arctic subpopulations to 0.199 between Baffin Island and Coastal Islands subpopulations, with mean pair-wise $F_{ST} = 0.125$. The Pacific subpopulation appeared most distinct by this measure, showing high F_{ST} estimates when compared to all other subpopulations (Table 2-1). A considerable amount of genetic variation (19.5%) was partitioned between subpopulations (AMOVA, $P < 0.0001$ after 9999 permutations) leaving 80.5% of the variance within subpopulations.

Subpopulation structure

The amount of substructure detected using the LD-pruned data set varied between subpopulations, but did not seem to depend on the number of individuals sampled or the area of land a given subpopulation inhabits. For the Forest subpopulation, which encompassed the largest area (see Fig. 2-2b) and the greatest number of individuals (n = 46 wolves), there appeared to be $K = 3$ subclusters. Two main genetic subclusters were aligned along an east/west gradient (Fig. 2-4a), with an additional third subcluster comprising the two wolves from Riding Mountain National Park. In the Arctic subpopulation (n = 9) STRUCTURE recovered two genetic subclusters (Fig. 2-4b). The wolves from Victoria and Banks Islands near the mainland formed one subcluster, and the wolves up in the high arctic, from Devon and Ellesmere Islands were admixed with a second genetic subcluster. The Pacific subpopulation (n = 8) was best represented at $K = 2$, with the two genetic subclusters aligned along an east/west split (Fig. 2-4c). Wolves near the coast formed one cluster, while wolves farther from the mainland were primarily composed of the second cluster. Neither the Baffin Island (n = 8) nor the Atlantic subpopulation (n = 16) showed evidence for substructure past $K = 1$ (data not shown).

Isolation by distance

Across all individuals (n = 132) there was a significant correlation between geographic distance and D_{IBS} ($r = 0.595$; Mantel test P = 0.001 after 1000 permutations). This correlation was stronger among wolves located more than ~300km apart than it was among those separated by shorter geographic distances (Fig. 2-5a). When this dataset was reduced to 61 individuals that were typed for microsatellite loci, there was a similar correlation between geographic distance and D_{IBS} ($r = 0.623$; P = 0.001) that remained weaker at short geographic distances (Fig. 2-5b). However, the correlation observed using the microsatellite

data was considerably weaker (r = 0.397; P = 0.001) at all geographic distances (Fig. 2-5c).

I used partial Mantel tests to separate the effects of subpopulation designation and geographic distance using the subset of 87 wolves that assigned strongly to subpopulations. The correlation between geographic distance and D_{IBS} after correcting for subpopulation designation was moderate (partial Mantel test; r = 0.438, P = 0.001). Similarly, there was a moderate correlation between subpopulation assignment and D_{IBS} after controlling for geographic distance (r = 0.463, P = 0.001).

Patterns of IBD within subpopulations were variable. There was a significant correlation between geographic distance and D_{IBS} in the Forest subpopulation (Mantel test; r = 0.594; P = 0.001), the Atlantic subpopulation (r = 0.529; P = 0.001) and the Arctic subpopulation, even after correcting for water barriers (partial Mantel test; r = 0.381; P = 0.012). However, there was not a significant correlation between geographic distance and D_{IBS} in the Baffin Island subpopulation (Mantel test; r = 0.039; P = 0.403) or in the Pacific subpopulation after controlling for water barriers (partial Mantel test; r = 0.073; P = 0.318).

Comparison of SNP data sets and microsatellites for determining population structure

For STRUCTURE runs with the LD-pruned SNP set, the 61-individual data set appeared to asymptote in Ln P (D|K) at K = 3 (Fig. 2-6), so this value was used to assess subsets of SNPs and microsatellite data. The clusters outlined were quite similar to the Arctic, Forest, and Baffin Island clusters recovered from the full data set (Fig. 2-7). The most noticeable difference was that the individuals on the east coast of the mainland assigned strongly to the Forest cluster, rather than a distinct Atlantic cluster. At K = 4, there was some distinction between these Atlantic samples and the rest of the Forest individuals, but the Atlantic individuals were still highly admixed (data not shown).

Mean discordance decreased with increasing number of SNPs, and the variation in mean discordance was best explained with an inverse-power best-fit line in both the full 132-individual data set ($r = 0.994$; Fig. 2-8a) and the subset of 61 microsatellite-typed wolves ($r = 0.992$; Fig. 2-8b). Clusteredness on the other hand did not follow a simple pattern. In the complete 132-wolf data set, mean clusteredness initially increased quickly with increasing number of SNPs before peaking in subsets of 500 SNPs, after which mean clusteredness decreased gradually as data subsets increased further in size (Fig. 2-8c). A similar pattern was observed in the 61-individual data set, but mean clusteredness peaked in subsets of 98 SNPs (Fig. 2-8d). Variance in both measurements tended to decrease as the number of SNPs in a subset increased.

Across both data sets, there was a significant negative correlation between the mean discordance of an individual and the best-estimate clusteredness in marker sets with > 56 SNPs ($P < 3.55 * 10^{-7}$). The strength of the correlation increased between subsets of 98 ($r = 0.512$ for 132 individuals; $r = 0.505$ for 61 individuals) and 500 SNPs ($r = 0.906$; $r = 0.941$), and then decreased in the 1000- and 5000-SNP subsets.

In the discordance measurement, the score for the 14-microsatellite data set fell in between the means of the 56-SNP and 140-SNP subsets, and in clusteredness the value was in between the means of the 56-SNP and 98-SNP subsets. Similarly, there was a negative correlation between discordance in the microsatellite data and best-estimate clusteredness ($r = 0.423$; $P = 1.67 * 10^{-3}$), but the correlation was weaker than in all SNP subsets with > 56 markers.

Discussion

Through the use of SNP markers I have found significant genetic structure in the North American grey wolf population with evidence for five subpopulations, but a large number of individuals (n = 45) are highly admixed between two or more subpopulations. I observed moderate genetic differentiation

between these subpopulations and found a strong pattern of isolation by distance across the entire population and within several subpopulations. Finally, I found high discordance between genetic cluster assignments when using small data sets and the entire set of SNPs, and in particular found evidence that highly admixed individuals show particularly high discordance when assigned with small data sets. The implications of the different results are discussed separately below.

Population structure

While I inferred that there are five main subpopulations of wolves across the study range, Carmichael *et al.* (2007), who sampled wolves across a range that was almost identical, found evidence for eight subpopulations of wolves. The contrasting pattern I observed has biological explanations, but there are also statistical reasons explaining why the pattern is not the same between these studies, and I will address these first.

Most of the subpopulations I recovered using SNP data were also recovered by Carmichael *et al.* (2007): Pacific, Arctic, Baffin Island, and Atlantic subpopulations. However, Carmichael *et al.* (2007) also detected two distinct mainland tundra subpopulations, as well as two forest subpopulations. My results on the other hand showed that mainland tundra wolves (wolves northeast of the tree line, visible in Fig. 2-2a) are highly admixed, containing large admixture proportions of both Forest and Arctic subpopulations (Fig. 2-2b), and that the boreal forest contains a continuous subpopulation of wolves. A discrete mainland tundra cluster does not appear in STRUCTURE runs until $K = 7$, and even then individuals in this cluster are still highly admixed, as are many Forest wolves (data not shown). Additionally, there is no evidence from the PCA of a mainland tundra subpopulation or two separate Forest subpopulations. There are a couple possible explanations for the reduced number of subpopulations. Carmichael *et al.*'s (2007) sample set was much larger (~2000 individuals) than the sample set used in this study (132 samples), but the distribution of their samples was similar

to the distribution of my samples. However, the sampling distribution in this study is more even than that of Carmichael *et al.* (2007), in which there were numerous individual capture locations where tens of animals were sampled. In this study there are only six locations with multiple individuals sampled, each with ≤ 5 individuals. Therefore, it seems likely that the discrepancy in population structure may be caused by uneven sampling and/or localized patterns of spatial autocorrelation (see Schwartz and McKelvey 2009) in Carmichael *et al.*'s (2007) study, rather a lack of samples in this study.

Additionally, the discrepancy in structure detected could be the result of the markers used (microsatellites vs. SNPs) and their ascertainment. Microsatellites are highly variable and the specific microsatellites used for genotyping are usually chosen because they exhibit a large number of alleles, and therefore over-represent genomic diversity (Brandström and Ellegren 2008). On the other hand, all SNPs in this data set have only two alleles, and the only SNPs removed from the data set were those with a minor allele frequency < 5%. Additionally, there is much greater genomic coverage offered by these SNPs, and so it is probable that this SNP data set more accurately depicts genome-wide diversity than microsatellites. Thus, the greater structure of North American grey wolves suggested by Carmichael *et al.* (2007) may be due to an over-estimation of genetic diversity as well as uneven sampling.

A new finding from this study is that the wolves found on the mainland arctic tundra (northeast of the treeline in Fig. 2-2a) do not form a distinct subpopulation, but are highly admixed between Forest and Arctic subpopulations. This admixture can be explained by wolves in the arctic tundra (both on the mainland and the islands) relying primarily on caribou for food (Parker 1973; Walton *et al.* 2001). Carmichael *et al.* (2007) and Musiani *et al.* (2007) found a strong correlation between population structure in wolves and primary prey type: mainland tundra wolves that prey upon barren-ground caribou rather than non-migratory prey [such as elk (*Cervus canadensis*), deer (*Odocoileus virginianus*; *O. hemionus*), or moose (*Alces alces*) found in the boreal forest] formed distinct

subpopulations. Barren-ground caribou herds have been observed migrating as far as 1200km (Anand-Wheeler 2002), and mainland herds move from tundra to forest and back again on an annual basis, giving birth on the tundra during summer and retreating south of the treeline for the winter (Hall 1989). Grey wolves from the tundra have long been thought to follow them during their semi-annual migrations (Walton *et al.* 2001). This has recently been confirmed using satellite telemetry (Musiani *et al.* 2007), with wolves observed following the caribou into the forest during the winter, and then back up onto the arctic tundra during the summer.

Barren-ground caribou inhabit the mainland tundra (northeast of the treeline) as well as adjacent islands including Baffin Island and the southern half of Victoria Island, and have been observed migrating between the islands and the mainland (Hall 1989; Anand-Wheeler 2002). Carmichael *et al.* (2001; 2008) inferred that the annual migration of the Dolphin-Union caribou herd between Victoria Island and the mainland increased gene flow between Victoria Island and mainland tundra wolves. It is therefore not surprising that the wolves of the mainland tundra appear highly admixed, as their main source of food travels between the ranges of three wolf subpopulations identified in this study (Arctic, Forest, and Baffin Island). Since wolves in this area follow their prey's annual migration, many opportunities exist for mainland tundra wolves to mate with wolves from the boreal forest as well as the arctic islands. In addition, wolves from more northern islands show less admixture with both Forest and Baffin Island subpopulations than do mainland tundra wolves. These islands (Devon, Ellesmere, and Banks) are inhabited by Peary caribou (*Rangifer tarandus pearyi*), rather than barren-ground caribou. Peary caribou undergo annual intra-island migration, but there have been few observations of inter-island migration and they do not inhabit the mainland (Hall 1989). As such, the wolves on these islands are likely more sedentary than wolves that prey on barren-ground caribou, and do not likely cross subpopulation boundaries on a yearly basis. Forest and Arctic subpopulations also had the lowest pair-wise F_{ST} across the entire population, despite occurring in different biomes and having a large water barrier between

them. Only two pair-wise subpopulation comparisons within biomes were possible, and both groupings (Forest vs. Atlantic and Baffin Island vs. Arctic) exhibited a lower pair-wise F_{ST} (0.072 and 0.116 respectively) than the mean pair-wise F_{ST} (0.125), but greater F_{ST} than between Arctic and Forest subpopulations. This low genetic differentiation is likely a result of high gene flow between Arctic and Forest subpopulations mediated by the annual migration of the highly admixed mainland tundra wolves following their prey.

Surprisingly, the wolves on Baffin Island showed only a small amount of admixture with other subpopulations, despite the near-ubiquitous presence of barren-ground caribou (Figure 2-2a). This is likely because migration of the caribou on Baffin Island largely occurs within the island. Only small numbers of caribou from the northwest of the island have been observed travelling to the mainland (Hall 1989), possibly because there is only a short stretch of water across which Baffin Island is adjacent to the mainland (the Fury and Hecla Strait). There is likely little contact between mainland wolves and the majority of the Baffin Island subpopulation as few wolves would travel between Baffin Island and the mainland while following caribou. Corroborating this, Carmichael *et al.* (2008) found greater genetic distance between north Baffin Island and mainland wolf subpopulations than between south Baffin Island and mainland wolf subpopulations. This explains why the Baffin Island subpopulation remains largely internally homogeneous and appears distinct from the Arctic subpopulation despite their close proximity and similar habitat.

Carmichael *et al.* (2007) were unconvinced of the presence of a legitimate subpopulation of wolves off the coast of the Alaskan panhandle in the Alexander Archipelago; contrarily, I have found strong evidence that this subpopulation is real and not simply the by-product of having over-represented a small geographic area as previously suggested. First, this subpopulation of wolves appeared distinct in STRUCTURE analyses when $K = 2$, suggesting these wolves are more genetically distinct from the rest of the samples than any other genetic cluster. This was confirmed with the pair-wise F_{ST} analyses. Second, this group of wolves

appeared distinct in PCA analysis, and occupies a significant portion of the first principal component. Unlike STRUCTURE, PCA does not group individuals into breeding subpopulations, and the results are not biased by sampling density or inaccurate allele frequencies. Last, I genotyped two individuals that assigned > 50% to this subpopulation from Vancouver Island, which is over 400km south of the Alexander Archipelago and separated from it by water barriers, again indicating that this genetic cluster is not the result of over-sampling. Collectively, this suggests that there is a strong differentiation between Pacific wolves and the rest of the subpopulations at a continent-wide scale. This finding agrees with Weckworth *et al.* (2005), who found significant distinction between wolves on coastal Alaskan islands and mainland forest wolves when assessed using 11 microsatellites. Similarly, Muñoz-Fuentes *et al.* (2009) found strong mitochondrial genetic differentiation between wolves found on opposite sides of the Coast Mountains, which separate interior and coastal British Columbia. Given that the climates and available prey types on opposite sides of the Coast Mountains are distinct (Muñoz-Fuentes *et al.* 2009), my result supports the hypothesis that large-scale grey wolf population structure is strongly influenced by ecological factors (Pilot *et al.* 2006; Carmichael *et al*. 2007).

Finally, Carmichael *et al.* (2007) found that North American grey wolf population structure based on microsatellite loci did not match the distribution of morphological subspecies described by Nowak (1995). The different structure proposed here supports this conclusion with two exceptions. The distribution of *C. l. occidentalis* proposed by Nowak (2003; Fig. 2-9) is very similar to the distribution of the Forest subpopulation described here (Fig. 2-2b) and the distribution of *C. l. arctos* closely matches that of the Arctic subpopulation. This correlation, however, is contradicted by the finding that the distribution of *C. l. nubilus* contains Pacific, Atlantic and Baffin subpopulations, as well as the highly admixed individuals of the mainland tundra. These subpopulations all show substantial allele frequency differentiation (Table 2-1), further refuting their status as a single subspecies. Thus, while SNP data in this study support the *C. l.*

occidentalis and *C. l. arctos* subspecies classification, my results do not support *C. l. nubilus* as a genetically distinct subspecies.

Genetic differentiation across subpopulations

Wolf subpopulations across northern North America appear moderately differentiated, with global F_{ST} = 0.131 and mean pair-wise subpopulation F_{ST} = 0.125. However, these results are upwardly biased, since highly admixed individuals were not included in this analysis. In comparison, Roy *et al.* (1994) estimated a mean F_{ST} = 0.167 across seven North American wolf subpopulations genotyped at ten microsatellite loci. This estimate is higher than the mean F_{ST} = 0.131 I observed, but is remarkably close to the estimates from the SNP data set. While my samples are spread out across almost the entirety of Canada and Alaska, Roy *et al.* (1994) sampled wolves from discrete areas, which will further skew F_{ST} values upwards relative to the sampling scheme in my study, accounting for the small discrepancy between our results.

In contrast, Carmichael *et al.* (2001) found a wide range of pair-wise F_{ST} values (0.009 – 0.188) between subpopulation pairs in the Canadian northwest, sampled both on the islands and on the mainland, suggesting high genetic differentiation between arctic islands wolves and mainland wolves. In contrast to this study, the scale of their analysis was much smaller, and subpopulation delineation was based solely on capture location, rather than results from clustering algorithms. Interestingly, all wolves sampled in their study were within or between the ranges of the Arctic and Forest subpopulations identified in this study, which have the lowest pair-wise F_{ST} of all subpopulation pairs (F_{ST} = 0.065). The high F_{ST} values estimated by Carmichael *et al.* (2001) could be the result of an over-estimation of genomic diversity (Brandström and Ellegren 2008) or sampling error (Ruzzante 1998), both of which can result from the use of microsatellite markers. Similarly, the very low values observed between other

subpopulation pairs are likely an artifact of subpopulation assignment because the groupings used do not necessarily represent meaningful population structure.

More recently, Aspi *et al.* (2009) estimated $\phi_{ST} = 0.151$ between a pair of Russian and Finnish wolf populations separated by < 200km. This value is comparable to F_{ST} values between Pacific and Atlantic subpopulations (0.166), or between Atlantic and Baffin Island subpopulations (0.144), which is surprising, since these subpopulation pairs are separated by > 3000km of over-land distance and occupy different biomes. Aspi *et al.* (2009) explain their high genetic differentiation by speculating on possible human-caused barriers in the form of Soviet-era fences and contemporary hunting. My result of comparable F_{ST} values between subpopulation pairs that are considerably farther apart, separated by intervening wolf subpopulations, and occupying different biomes suggest that the genetic differentiation calculated by Aspi *et al.* (2009) is also overestimated. This is confirmed by Aspi *et al.*'s (2009) AMOVA analysis, which showed that ~15% of genetic variation occurred between the Russian and Finnish populations. This is less than the variation occurring between North American wolf subpopulations (~20%), and again indicates that their estimation of ϕ_{ST} is likely exaggerated.

Because previous studies have used microsatellite markers, sampling bias and an over-estimation of genetic diversity may have skewed F_{ST} values, so that some populations of wolves have looked more genetically distinct than they truly are. The use of a more even sampling scheme, combined with a marker more appropriate for assessing genetic diversity and a more sophisticated method of subpopulation delineation has provided estimates that are significantly lower than in some previous studies, suggesting that grey wolf subpopulations are more similar than previously documented.

Subpopulation structure

From the investigation of subpopulation structure it is clear that this large SNP data set is capable of showing fine-scale resolution in the data. However,

substructure was not evident in all subpopulations, and may depend on the ecology and geographical features of the subpopulation in question. This can be illustrated with the Forest subpopulation. Containing the largest number of individuals over the greatest physical area, I hypothesized that there would be a high amount of substructure within this subpopulation. The east-west divide observed matches the findings of Carmichael *et al.* (2007) who found two distinct Forest subpopulations with an east/west divide, and is not surprising given the scale of the subpopulation, with the two most distant individuals being separated by > 3500km. More surprising is that the additional cluster detected by STRUCTURE was only found in large proportions in two wolves from Riding Mountain National Park, and accounts for < 60% of their assignment. These wolves show little pair-wise genetic differentiation with $D_{IBS} = 0.149$, approximately half of the mean D_{IBS} (0.280) for this subpopulation. Thus, it seems that STRUCTURE identified individuals that may be from the same pack, and do not represent additional geographic structure within this subpopulation. Further, based on analyses including coyote and domestic dog samples, these two individuals appear to be partially admixed with coyotes (vonHoldt *et al.* in prep.), which would also account for their strong distinction within this subpopulation. Overall, the Forest subpopulation of wolves is relatively homogenous, as the main pattern of differentiation is clinal rather than discrete. Wolves have been observed migrating very large distances (> 800km, Fritts 1983; > 1000km, Wabakken *et al.* 2007), and the habitat across this area is fairly homogenous, with most of the subpopulation occupying the boreal forest biome. Because the habitat occupied by these wolves is continuous and because no major barriers to gene flow exist within this subpopulation it is logical that such a highly mobile animal showed only a gradient in genetic differentiation across this large area.

Both Pacific and Arctic subpopulations showed substructure that can be attributed to water barriers. In the Pacific subpopulation, all individuals are geographically close, but individuals closer to the mainland formed a distinct genetic cluster, and individuals farther from the mainland assigned highly to a second cluster. Similarly, in the Arctic subpopulation, individuals closest to the

mainland (on Victoria and Banks islands) formed one genetic cluster, whereas those in the far north on Ellesmere and Devon islands showed admixture with a second cluster. These Arctic groups are much farther apart than those of the Pacific subpopulation, and yet the Pacific subpopulation showed stronger internal differentiation (see Fig. 2-4b, Fig. 2-4c). These patterns indicate that water barriers may reduce gene flow as suggested by Carmichael *et al.* (2001), but that the strength of water barriers varies between different areas. The water separating Arctic individuals is frozen for much of the year, so wolves would be free to travel between the nearer and farther islands, albeit across a very long distance. However, in the Alexander Archipelago where the Pacific wolves are located, the water does not freeze over the winter but is very cold year round and so may act as a strong barrier to gene flow despite the short geographic distance separating individuals. These findings highlight the importance of local geographic features to genetic differentiation between wolves, and emphasize that the effect of similar features may differ between regions.

Isolation by distance

I found evidence for IBD based on a significant correlation between D_{IBS} and Log_{10}-transformed geographic distance. This trend is weaker at distances < 300km, meaning that short-range distance has little effect on differentiation. This finding is consistent with wolves' ability to disperse long distances, and indicates that there is a high amount of gene flow occurring within the North American wolf population at relatively short distances. This suggests that while subpopulation differentiation occurs over large scales, it is unlikely to occur across small distances without some meaningful barrier to gene flow. A further implication is that there should be transition zones between subpopulations, containing genetically mixed individuals. This is confirmed by the large number of highly admixed individuals observed (n = 45; Figure 2-2a).

I found that geographic distance was significantly correlated with D_{IBS} even after controlling for population structure, in contrast to recent research indicating that geographic distance is not a strong factor underlying wolf genetic differentiation on a continent-wide scale (Pilot *et al.* 2006; Carmichael *et al.* 2007). Most analyses of IBD use population-based statistics such as Nei's (1972) D_S and F_{ST}. These analyses rely on estimated allele frequencies and averaged geographic distances and so may yield high noise when compared to individual-based analyses. However, Musiani *et al.* (2007) found no evidence of spatial autocorrelation across 14 microsatellite loci in wolves of the boreal forest and arctic tundra. Thus, the discrepancy may also be caused by the use of microsatellite markers in previous studies, rather than just the type of analysis *per se*. This is consistent with the microsatellite data across a common set of 61 wolves, which indicates that the stronger IBD observed in this study is (at least in part) a result of the markers used rather than some idiosyncrasy of the data set or the sampling scheme. The weak correlation observed is a result of the weak power of microsatellites to detect IBD, reflected in high noise across all geographic distances (Fig. 2-5c). Microsatellites generally have a high number of alleles (Vignal *et al.* 2002), and this will cause high noise and variation in allele-sharing distances when compared to SNPs, which only have two alleles. This was certainly the case in this study, as the number of alleles observed in each microsatellite marker ranged from 4-10, with a mean of 8.5. This high relative noise in microsatellite data was exacerbated by the fact that many more SNPs than microsatellites were genotyped, which resulted in more precise estimates of genetic distance between individuals when SNP data were used. Together, these differences may explain why IBD appears strong in this study but not in previous studies. Regardless of the relative importance of marker type vs. analysis type, it is likely that previous studies have underestimated the importance of geographic distance to large-scale population structure in grey wolves.

IBD was also observed within most subpopulations. The Pacific subpopulation did not show IBD, but this is unsurprising given that all individuals were < 300km apart. An anomaly is the Baffin Island subpopulation, which also

did not show evidence of IBD. There are no water barriers within this subpopulation, nor are the comparisons biased by a large number of individuals that are < 300km apart. The lack of IBD may be related to the movement of barren-ground caribou on Baffin Island. Because Baffin Island wolves follow barren-ground caribou as they migrate throughout the year, their recorded capture locations may not be meaningful representations of their home ranges, so a significant amount of noise in IBD plots is expected.

The presence of IBD is confirmed by the reduced structure I have observed compared to Carmichael *et al.* (2007). I have detected fewer genetic clusters (five vs. eight), and a large number of individuals (~1/3 of the samples) that are highly admixed between two or more clusters. The high genetic continuity across this population reflects the higher IBD detected using SNP data and confirms the utility of using high-resolution genetic data to examine population structure.

A concern when assessing population structure is that where there is significant IBD, clustering algorithms (such as STRUCTURE) can spuriously detect multiple subpopulations rather than indicating that there is a single breeding population, even if there truly is only one genetic cluster of individuals (Guillot *et al.* 2009). Despite the significant IBD observed in these wolves there is evidence that the results from STRUCTURE accurately represent population structure. The partial Mantel analyses showed a higher correlation between subpopulation assignment and D_{IBS} than between pair-wise geographic distance and D_{IBS}, despite the noise in the simplistic subpopulation-assignment distance matrix (i.e. because all distances are either "1" or "0"). This indicates that while geographic distance is highly influential, the genetic clusters determined by STRUCTURE are meaningful entities, and not simply arbitrarily selected groups of individuals.

Comparison of SNP data sets and microsatellites for determining population structure

As expected, I found that using a larger number of markers to determine population structure significantly improved assignment estimates (see Fig. 2-8a, 2-8b). By using the STRUCTURE assignments from the complete 26k data set as an approximation of correct subpopulation assignment, I evaluated the relative performance of STRUCTURE when run with subsets of SNP data, using a measure of discordance. Discordance was small and varied little across larger subsets of SNPs, but increased dramatically and varied widely across very small SNP subsets, indicating that smaller data sets may do a poor job of estimating total genomic diversity within individuals. This increase in discordance (and variation in discordance) with a decrease in markers can be interpreted as an increase in statistical noise caused by poor genomic coverage. This is conceptually similar to the effect of incomplete lineage sorting in phylogenetic analyses. Much as using only a small number of genes to infer a phylogeny can yield inaccurate or imprecise results, with conflicting phylogenies generated by different genes (e.g. Rokas *et al.* 2003), inferring population structure using a small number of markers yielded inaccurate results that varied substantially across data sets of the same size. By increasing the number of markers used, assignments improved asymptotically, getting decreasingly closer to the (assumed) correct set of assignments.

I found that mean discordance and the number of SNPs used to evaluate population structure appeared to have an inverse power law relationship. Thus, as the number of markers increased, the improvement in assignments relative to smaller data sets decreased. This was expected, as any increase in SNP markers brings the total information content closer to that in the complete 26k data set, and so the relative improvement decreases. In both 132-individual and 61-individual data sets I observed very low mean discordance when 5000 SNPs were used to evaluate population structure (0.034 and 0.019 respectively), and variance in this measure was very small. Mean discordance remains relatively low even when

only 500 SNPs (~2% of the complete data set) are used (0.106 and 0.064 respectively). Running STRUCTURE with < 500 SNPs however yielded substantially more discordant results; with 140 SNPs, discordance increased dramatically with means of $d = 0.165$ and 0.131, and a single subset of SNPs showing $d = 0.256$ in the 132-individual data set. This indicates that 500 SNPs may approximate a reasonable balance between expense (in terms of genotyping) and accuracy in assignment.

Surprisingly, clusteredness did not simply increase with the number of markers across the range, and that after reaching a threshold, actually decreased with increasing markers (see Fig. 2-8c, 2-8d). That is, in small SNP subsets, individuals appeared to be highly admixed. This admixture initially decreased with increasing data but eventually stopped, and then admixture slowly increased with the number of markers used. This indicates that small SNP subsets lack the power to strongly assign individuals to the appropriate genetic cluster, but slightly larger subsets may show greater genetic differentiation between individuals than is accurate. These SNP subsets have strong power to discriminate between genetic clusters but have relatively poor resolution compared to the complete set of markers, and are less able to identify admixture. This finding was confirmed by the observed negative correlation between individuals' best-estimate clusteredness and mean discordance across SNP subsets of the same size. A significant negative correlation was found in all but the smallest SNP subsets and was strongest in SNP subsets containing 500 SNPs across both 132- and 61-individual data sets. This means that discordance is driven primarily by individuals with mixed genetic ancestry and suggests that with all but the largest SNP data sets admixed individuals will be highly mis-assigned. Indeed, across 500-SNP subsets, some highly admixed individuals from the 132-individual data set showed mean discordance values in excess of 0.3 (data not shown), which is approximately three times the mean $d = 10.6$ across all individuals, and greater than the mean discordance observed across 56-SNP subsets. This result has implications for any study where the accurate estimation of admixture is of critical importance, such as the detection of hybrid zones for conservation purposes (Allendorf *et al.* 2001),

assessing gene flow between sympatric species (e.g. Ochieng *et al.* 2008), or describing the evolution and spread of antibiotic resistance in bacteria (Hanage *et al.* 2009).

To my knowledge, this is the first study to assess the performance of data sets composed of differing numbers of SNPs for the purposes of determining population structure. However, the number of SNPs necessary to estimate inbreeding and relatedness in zebra finches has recently been evaluated (Santure *et al.* 2010). Santure *et al.* (2010) observed a strong correlation between relatedness estimates based on 771 SNPs and pedigree data, but found that little improvement in genetic relatedness estimates was achieved by using subsets of their SNP data with > 500 markers. This is similar to the trend observed in this study. While I observed more accurate assignment and admixture using > 500 SNPs, the relative improvement in discordance was markedly less than the improvement observed by augmenting smaller data sets. The fact that 500 SNPs appears to approximate a threshold in both studies cannot be extrapolated to state that a 500 SNP data set will provide equally good results for all studies and species, but it does indicate that for many population genetics studies (where accurately assessing admixture is not highly important) genotyping individuals at several hundred SNP loci should yield sufficiently reliable results.

Something I was unable to explore is the effect that the number of samples, the sampling scheme, and the number of detected clusters had on the power of SNP subsets to accurately assigning individuals. I observed lower absolute discordance values in the 61-individual data set than the 132-individual data set, but this result was confounded by the fact that STRUCTURE was run at a different *K* value for each data set. Additionally, the clusteredness peaked at different sizes of SNP subsets across the different data sets, but this measurement too was confounded by *K*. Finally, the 61-individual data set was not randomly selected from the complete set of individuals, and so their distribution across the sampling range was more clumped than the full data set. This may in turn have led to lower detected levels of admixture, and thus lower discordance when

subsets of SNPs were used to assign individuals. A complete understanding of the interaction between all these factors would require increased sampling across this range, so that the influence of each factor could be evaluated separately. This means that despite the very strong correlation between mean discordance and number of markers in a SNP subset, it is not possible to estimate the absolute amount of discordance for a given number of markers for other data sets.

In both discordance and clusteredness, the results from the microsatellite data set had values that fell in between the mean values of the 56-SNP and 140-SNP data subsets for the 61-individual data set. This suggests that ~4-10x as many SNPs as microsatellites are necessary to obtain a similar level of accuracy when assessing population structure. This inference is comparable to the results of Mariette *et al.* (2002), who estimated based on simulations that up to 10x as many bi-allelic AFLP markers as microsatellites are required to estimate genomic diversity. Unfortunately, because the variance in discordance and clusteredness between SNP subsets with < 500 markers was fairly large, this estimate is imprecise. A more precise estimate of the equivalent number of SNPs per microsatellite marker will require genotyping individuals at a greater number of microsatellite loci so that multiple sets of microsatellites can be analyzed in order to determine the variance in their discordance and clusteredness. Additionally, there will necessarily be variance between any comparison of SNPs and microsatellites due to the wide range of the number of alleles observed in different microsatellites, and it is unclear what influence ascertainment scheme may have. Despite these caveats, my result should provide a useful comparison point for researchers considering using SNP markers rather than microsatellites for population genetic analysis.

Conclusions

In the course of this chapter, I have re-assessed the population structure of North American grey wolves, and evaluated a large set of SNP markers for their

use in population genetic analyses commonly used to evaluate wild species. While the analyses performed here are by no means exhaustive from a population genetics perspective, they provide significant insight into the behaviour of large sets of markers when subjected to traditional population genetic analyses.

Based on SNP data, North American grey wolves exhibit significantly higher isolation by distance than reported in previous studies. This finding suggests that previous assessments of IBD may lack power due to the relatively low resolution and high noise of smaller marker sets. Significant IBD has resulted in reduced genetic structure compared to microsatellite-based analysis, with fewer subpopulations detected and a high number of admixed individuals. Despite that, this study confirms the presence of several previously inferred subpopulations, and verifies the genetic distinction of a subpopulation of wolves found on islands off the Pacific coast. Notably, I have found that prey-specialization on migratory barren-ground caribou in the mainland tundra appears to have caused wolves across this large region to become highly genetically admixed rather than genetically distinct. Finally, I found that the structure observed here only partially agrees with the subspecies designations of Nowak (1995), indicating that subspecies designations based on skull morphology do not necessarily reflect underlying genetic differentiation.

There is high resolution in the genetic data, and I was able to detect genetic structure at a subpopulation level to the extent that I identified genetic subclustering in a subpopulation of only eight individuals. However, not all subpopulations show such distinct substructure, and two of the five detected subpopulations show no substructure at all. This highlights the idiosyncratic nature of population structure, and indicates that genetic structure may be highly contingent on local ecological and geographical features.

Finally, I have found high variation in subpopulation assignments based on marker sets containing fewer than 500 SNPs. While there are diminishing returns to increasing the number of markers used for assignment, I find that the use of relatively small sets of markers (as is common in population genetic

studies) may yield assignments with a high amount of inaccuracy, especially in populations with many admixed individuals. This finding showcases the need for continuing development of genetic markers for non-model organisms, because the potential to mis-characterize genetic structure when using < 500 SNPs appears to be significant. Additionally, even if the mean assignment is relatively accurate, there is still high potential for imprecision in individual assignments, and high inaccuracy in the assignment of admixed individuals.

Table 2-1. Mean pair-wise F_{ST} comparisons between five subpopulations of wolves, calculated across 26 221 SNPs.

subpopulation	Arctic	Atlantic	Baffin	Forest	Pacific
Arctic	0.000				
Atlantic	0.110	0.000			
Baffin	0.116	0.144	0.000		
Forest	0.065	0.072	0.103	0.000	
Pacific	0.167	0.166	0.199	0.111	0.000

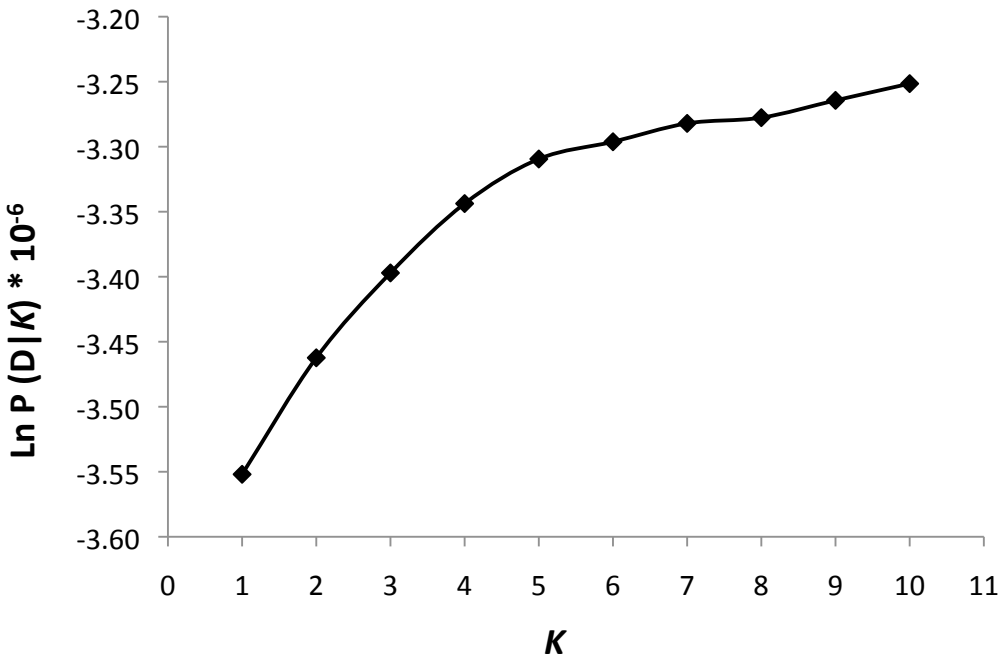


Figure 2-1. Number of clusters (*K*) vs. likelihood [Ln P (D|*K*)] for STRUCTURE analyses of 132 grey wolves using 26 221 SNPs with the admixture model. Each point represents the highest value of Ln P (D|*K*) obtained from across five replicate STRUCTURE runs with 20 000 burn-in cycles followed by 5000 sampling iterations.

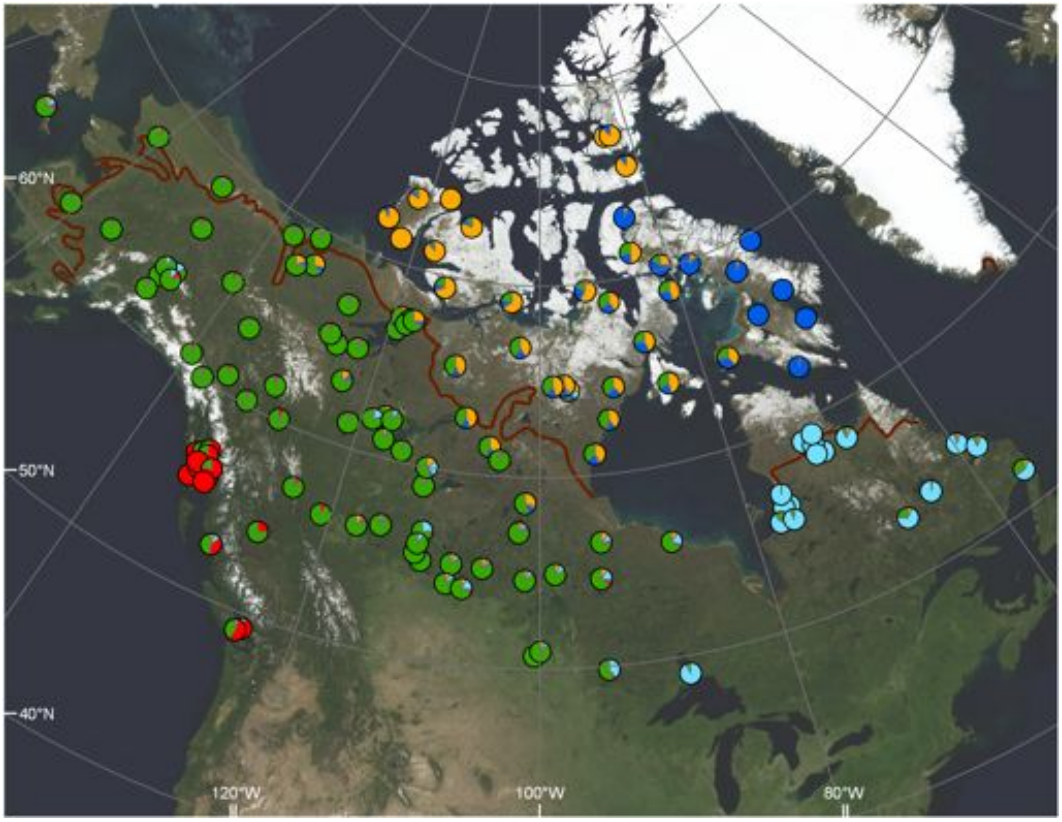


Figure 2-2a. Capture location and assignment of 132 grey wolves from Canada and Alaska. Each individual is represented with a pie chart, showing their admixture proportions for each of five genetic clusters determined from STRUCTURE analysis of 26 221 SNPs. The tree line, separating boreal forest and arctic tundra biomes, is shown in brown. Individuals from the same capture location were slightly displaced to aid visualisation.

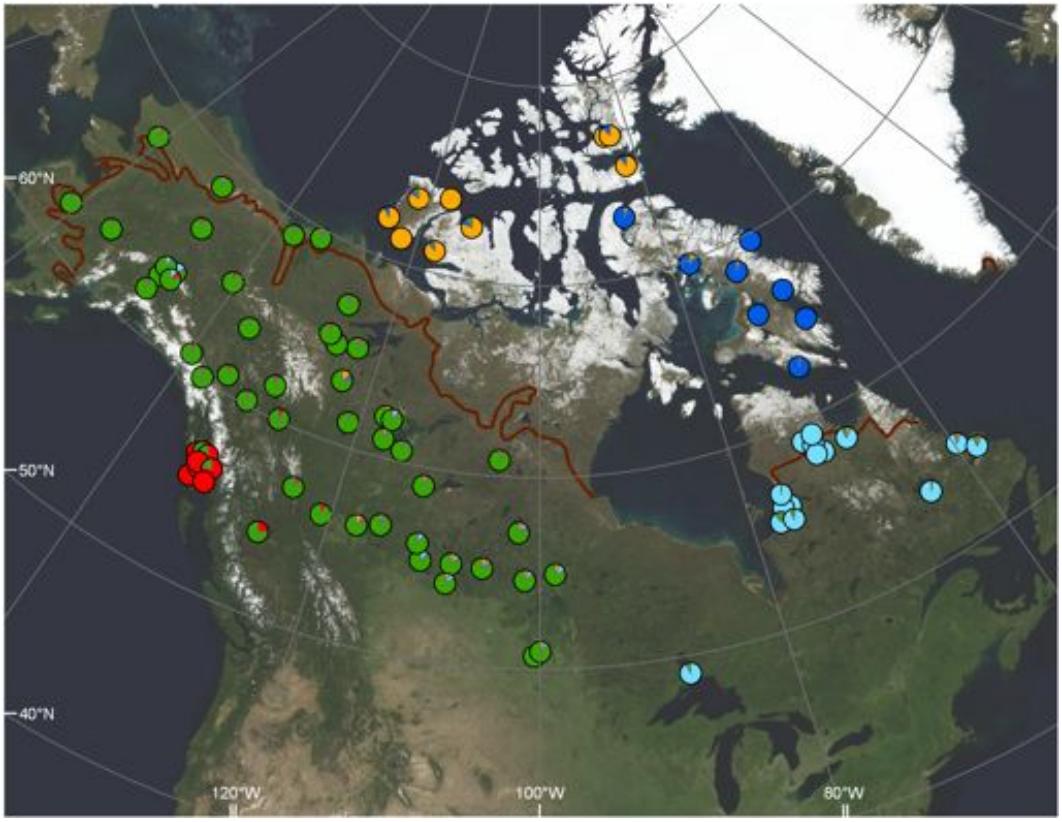


Figure 2-2b. Capture location and assignment of 87 grey wolves from Canada and Alaska highly assigned to one of five subpopulations – Pacific (red), Forest (green), Arctic (orange), Baffin Island (dark blue) and Atlantic (light blue). Each individual is represented with a pie chart, showing their admixture proportions for each of five genetic clusters determined from STRUCTURE analysis of 26 221 SNPs. The tree line, separating boreal forest and arctic tundra biomes, is shown in brown. Individuals from the same capture location were slightly displaced to aid visualisation.

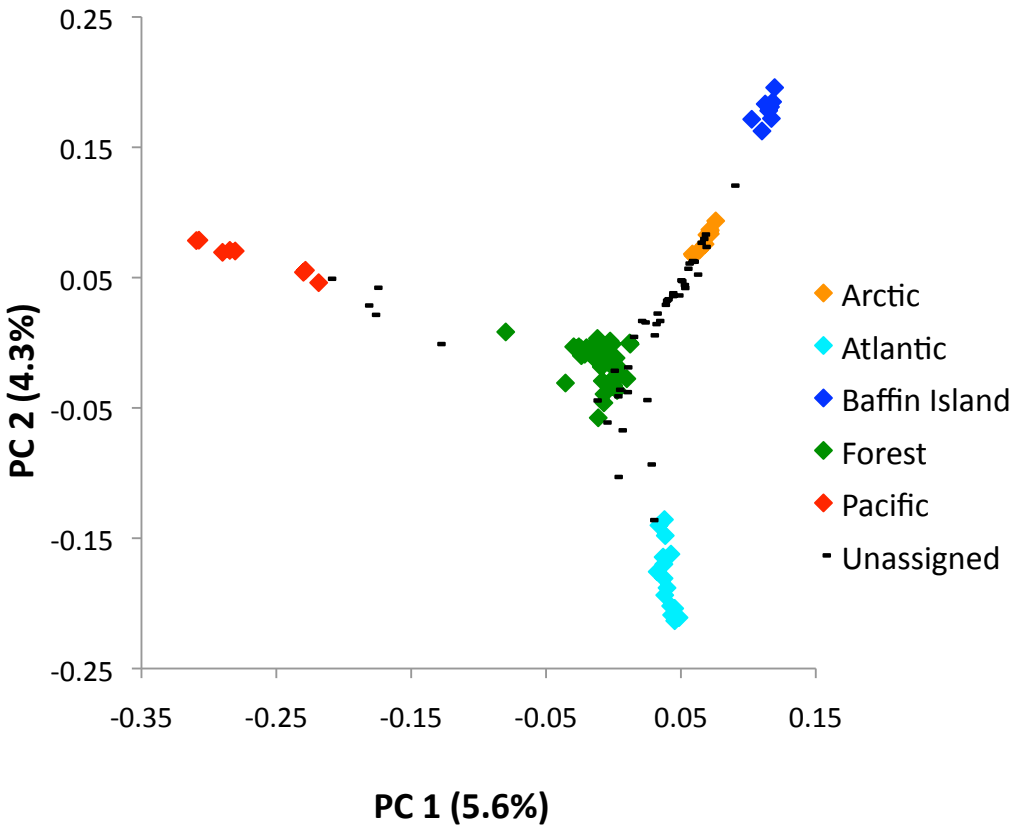


Figure 2-3. The first two principal components (PC 1, PC 2 respectively) from a PCA of 26 221 SNPs in 132 North American grey wolves. Subpopulation assignment determined using STRUCTURE analyses are shown. The amount (%) of genetic variation explained by each component is listed in parentheses.

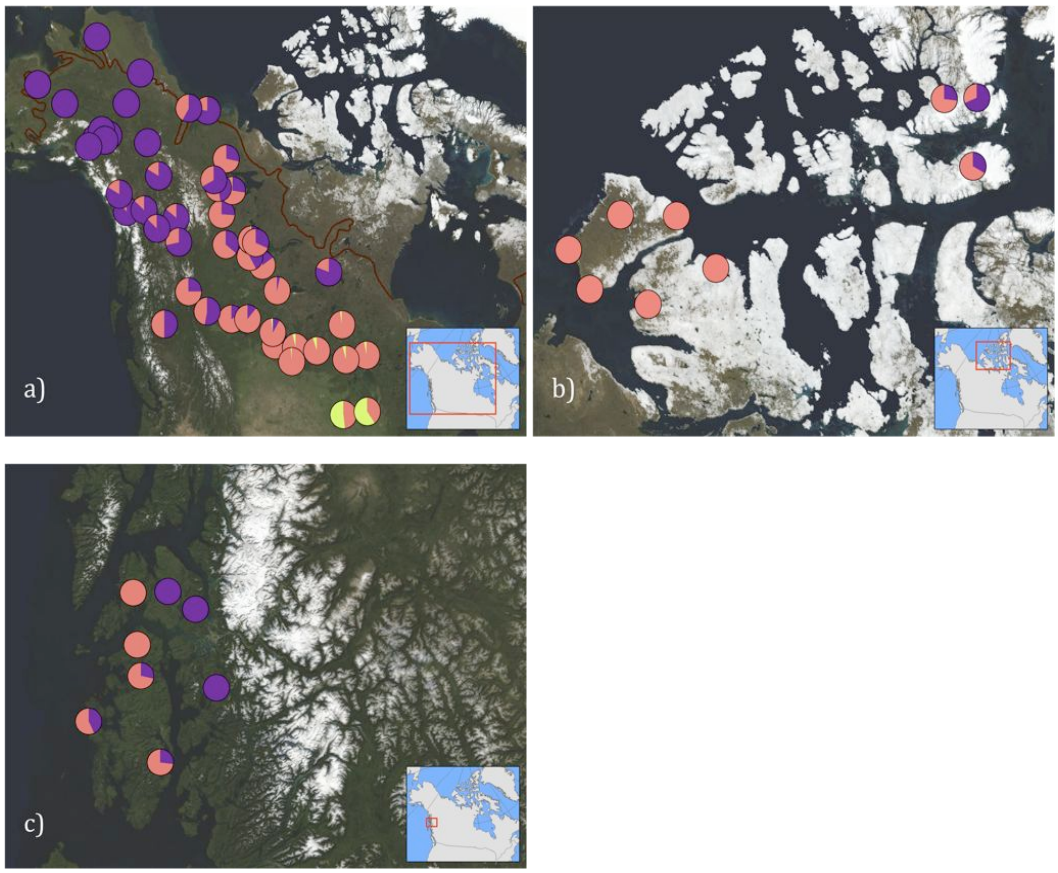


Figure 2-4. Capture location and assignment of a) 46 wolves from Canada and Alaska identified from the Forest subpopulation, b) nine wolves from Canada identified from the Arctic subpopulation and c) eight wolves from the Alexander Archipelago of southern Alaska identified from the Pacific subpopulation. Each individual is represented with a pie chart, showing their admixture proportions for each of three (a) or two (b, c) genetic subclusters as determined from STRUCTURE analysis of 26 221 SNPs. Four arctic islands are labelled in b): Victoria Island (VI), Banks Island (BI), Devon Island (DI), and Ellesmere Island (EI). The tree line, separating boreal forest and arctic tundra biomes, is shown in brown (a). Individuals from the same capture location were slightly displaced to aid visualisation.

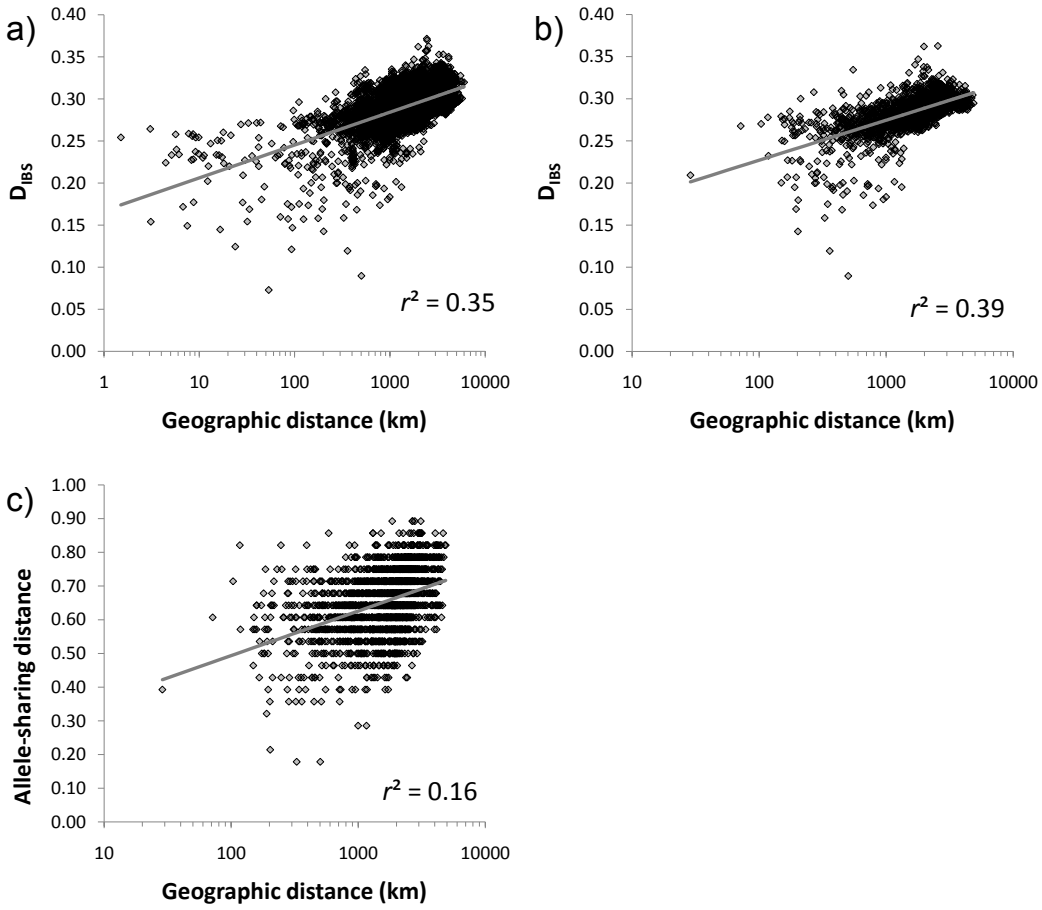


Figure 2-5. Correlation between pair-wise Log_{10}-transformed geographic distance and a) identity-by-state genetic distance between 132 wolves calculated from 26 221 SNPs, b) identity-by-state genetic distance between a subset of 61 wolves calculated from 26 221 SNPs and c) allele-sharing distances between a subset of 61 wolves calculated from 14 microsatellites. A linear best fit for each correlation is shown, and the proportion of the variation in genetic distance explained by Log_{10}-transformed geographic distance is reported (r^2).

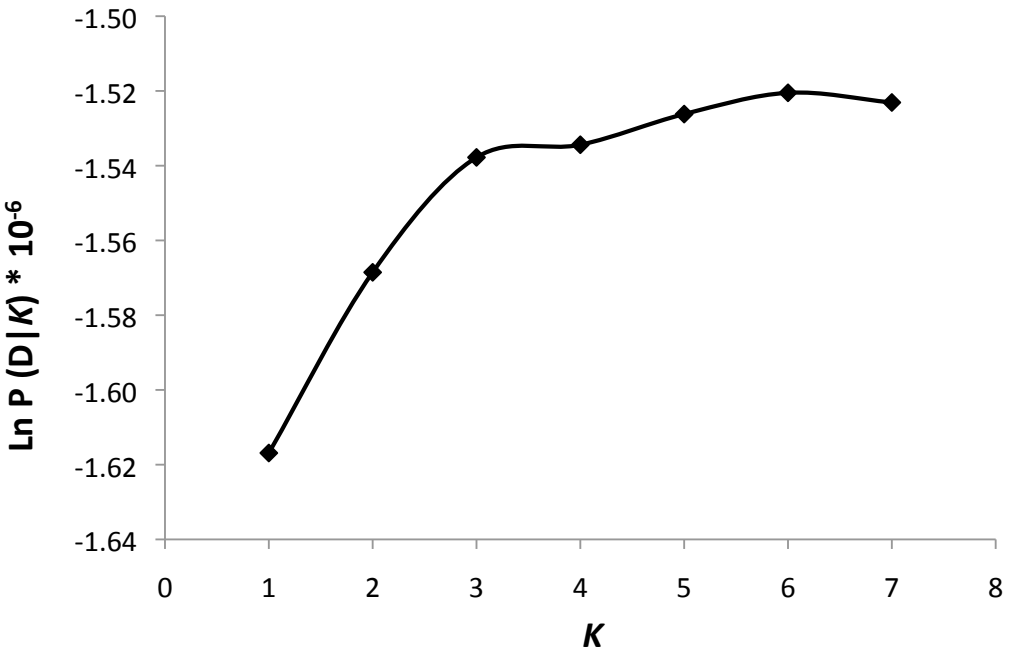


Figure 2-6. Number of clusters (*K*) vs. likelihood [Ln P (D|*K*)] for STRUCTURE analyses of 61 grey wolves [previously genotyped by Carmichael *et al.* (2007)] using 26 221 SNPs with the admixture model. Each point represents the highest value of Ln P (D|*K*) obtained from across five replicate STRUCTURE runs with 20 000 burn-in cycles followed by 10 000 sampling iterations.

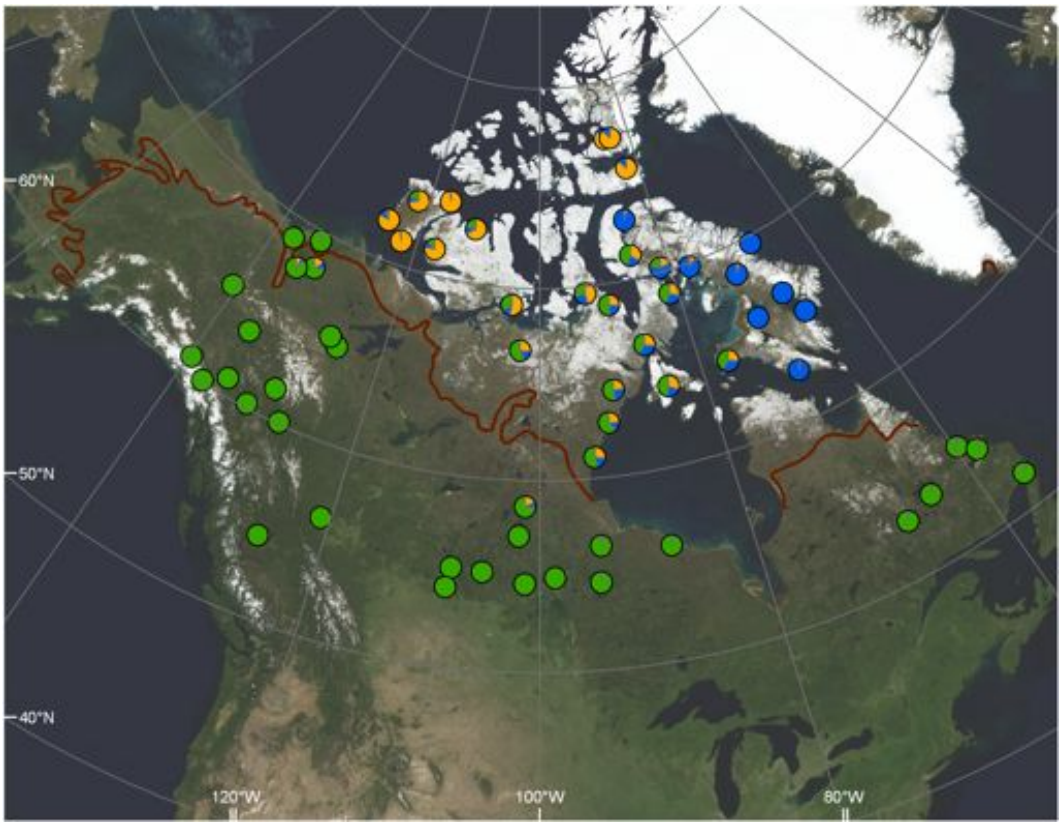
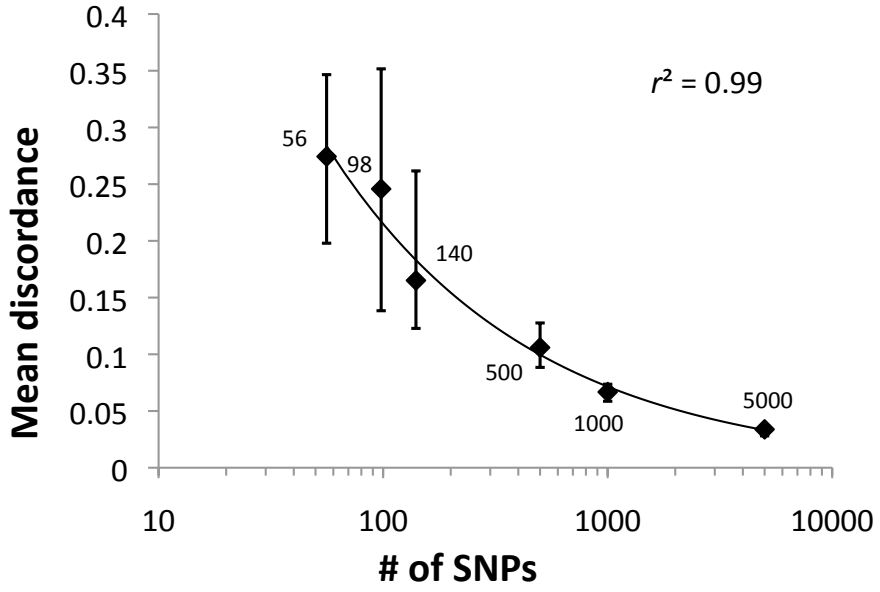


Figure 2-7. Capture location and assignment of 61 grey wolves previously genotyped by Carmichael *et al.* (2007). Each individual is represented with a pie chart, showing their admixture proportions for each of three genetic clusters determined from STRUCTURE analysis of 26 221 SNPs. The tree line, separating boreal forest and arctic tundra biomes, is shown in brown.

a)



b)

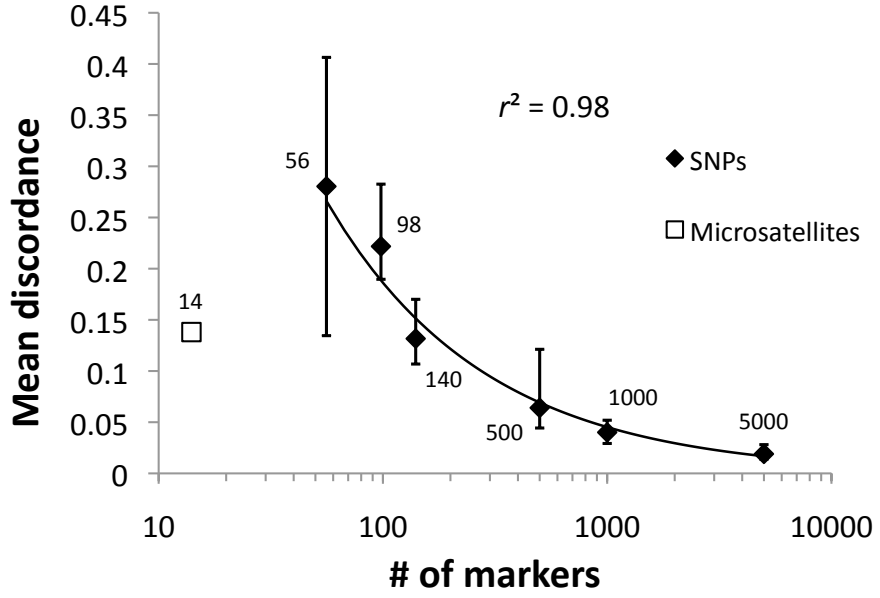
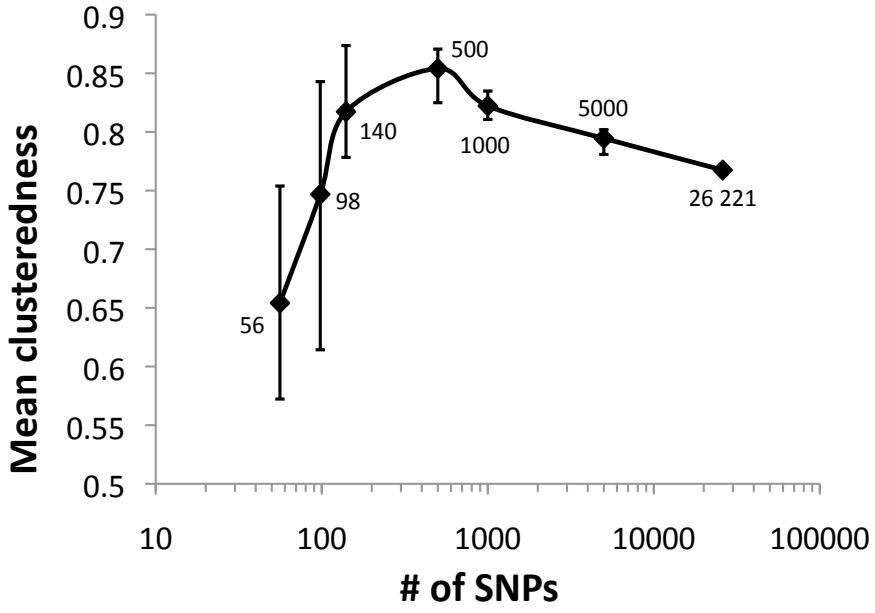


Figure 2-8a, b. Mean discordance plotted as a function of number of markers, comparing the number of markers used in randomly generated data sets to mean discordance of STRUCTURE assignment from assignment estimated from the complete 26 221 SNP data set for (a) all 132 wolves, and (b) the subset of 61 wolves genotyped by Carmichael et al (2007). An inverse-power regression line has been plotted in black, and the variation in mean discordance explained by this line is reported (r^2). Error bars show the range of mean discordance values from across ten replicate data subsets for each number of SNPs.

c)



d)

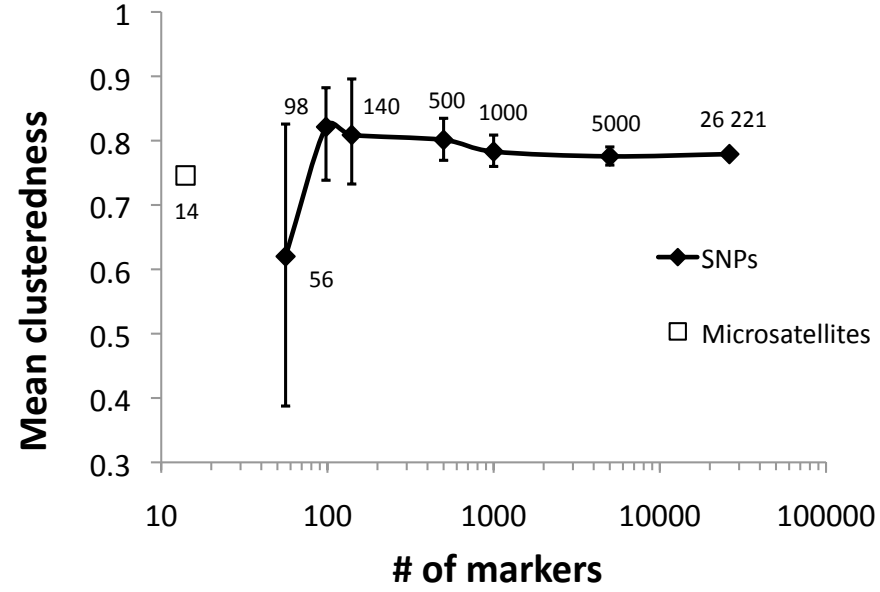


Figure 2-8c, d. Mean clusteredness plotted as a function of number of markers, comparing the number of markers used in randomly generated data sets to mean clusteredness of STRUCTURE assignments, as well as the clusteredness value observed when all 26 221 markers were used for (c) all 132 wolves, and (d) the subset of 61 wolves genotyped by Carmichael *et al.* (2007). Error bars show the range of mean clusteredness values observed from across ten replicate data subsets for each number of SNPs (not including complete 26 221 SNP data set).

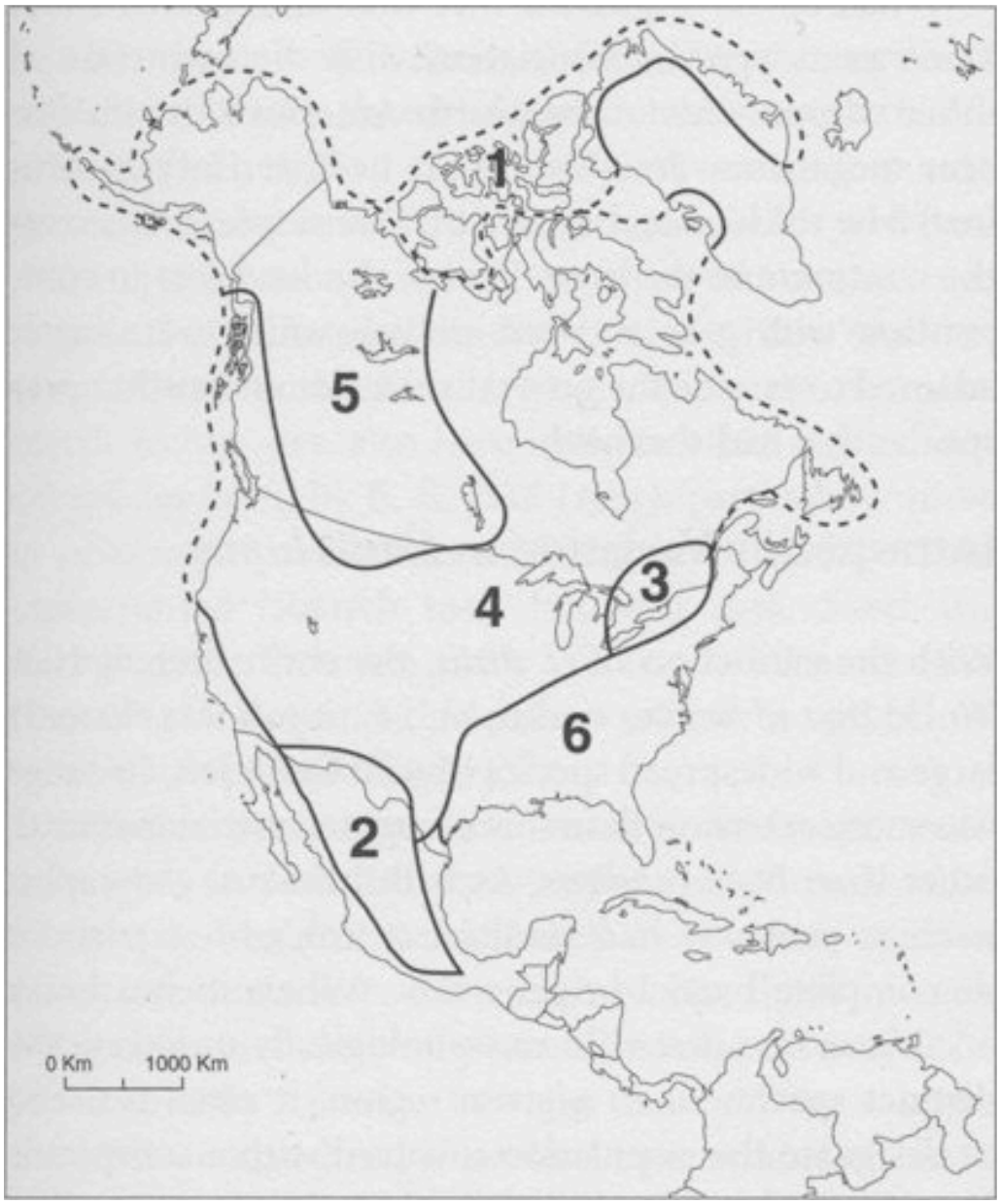


Figure 2-9. Original geographic distribution of wolves in North America, showing the two species and five subspecies recognized by Nowak (1995): 1, *C. l. arctos* (arctic wolf); 2, *C. l. baileyi* (Mexican wolf); 3, *C. l. lycaon* (eastern wolf); 4, *C. l. nubilus* (plains wolf); 5, *C. l. occidentalis* (northwestern wolf); 6, *C. rufus* (red wolf). Obtained from Nowak (2003), © 2003 by The University of Chicago.

Chapter 3
A genome scan to detect selected genes in North American grey wolves

Introduction

In the previous chapter I explored the uses of a large SNP data set for studying the population genetic structure of grey wolves. Population genetic analyses are routinely performed on wild populations using microsatellites and do not require the use of large panels of markers to get informative results. There are, however far fewer studies that have tried to find the genetic basis of phenotypic adaptations. Isolation of functional genetic variation has frequently involved mapping specific phenotypes onto quantitative trait loci (QTL; Slate *et al.* 2005). This approach requires not only a relatively large set of markers (usually either microsatellites or AFLPs in wild species) with which to map the variation in the genome, but also knowledge of the pedigree and specific individual phenotype measurements. Depending on the species of interest, phenotypes may be difficult to acquire and pedigree data may not be obtainable. Additionally, QTL analyses often lack the power (i.e. genomic coverage) necessary to pinpoint the specific gene(s) controlling a trait of interest, and thus require further research in order to isolate the functional gene(s).

With larger sets of markers and an increasing number of annotated genomes, it is now feasible to take a reverse approach to linking genetic and phenotypic data. By genotyping a number of individuals at many loci, it is possible to scan for signals of selection within the genome, locate genes near the selected loci, and relate these genes to previously observed phenotypic variation. Such genome scans have been used extensively to study selection in human populations (e.g. Nielsen *et al.* 2005; Sabeti *et al.* 2007; Pickrell *et al.* 2009), with findings including selection for genes controlling lactase synthesis (Nielsen *et al.*

2005) and skin pigmentation (Sabeti *et al.* 2007). Recently, this strategy has also been applied successfully to non-humans including cattle (Gautier *et al.* 2009) and domestic dogs (vonHoldt *et al.* 2010).

In this chapter, I applied a genome scan to a wild population of North American grey wolves to identify specific genes that may have undergone selection in this population. Of the various types of methods used to detect signals of selection (reviewed in Chapter 1), I used an empirical F_{ST}-outlier method. This method makes direct comparisons of the F_{ST}-values of all SNPs between groups of individuals in order to detect loci that show high allele frequency differentiation across populations. The idea behind this method [first proposed by Lewontin and Krakauer (1973)] is that any locus that has undergone directional selection and varies across groups should show higher allele frequency differentiation between these groups than neutral loci. By comparing the differentiation at many markers, outliers of this distribution can be interpreted as candidates of selection. An F_{ST}-outlier genome scan of this nature was applied successfully by Akey *et al.* (2002) to a set of ~28 000 SNPs genotyped across three human populations and has recently been used by vonHoldt *et al.* (2010) to detect signals of artificial selection in domestic dogs. By comparing allele frequency differences between genetic subpopulations of wolves (identified in Chapter 2, Fig. 2-2b), my goal was to locate genes that have undergone adaptive selection in North American grey wolves. Specifically, because ecological variables have been found to influence wolf population structure (Geffen *et al.* 2004; Pilot *et al.* 2006; Carmichael *et al.* 2007; Musiani *et al.* 2007), I expected that by looking for directional selection between subpopulations, I would find genes that could be linked to differing ecological selective pressures related to (for example) habitat, prey, and climate.

Methods

Detection of F_{ST} *outliers between subpopulations*

One way to identify markers that appear to be under directional selection is to select markers with allele frequencies that differ strongly between groups of individuals that are believed to be under different selective pressures. In order to detect candidates of directional selection in my data set, I grouped wolves into five subpopulations based on the results from Chapter 2 (Fig. 2-2b). Only wolves that had > 75% assignment to a single genetic cluster (n = 87) were analyzed. This way, I was confident in subpopulation assignment and avoided biasing allele frequencies with highly admixed individuals. Using a script written by J. Novembre, Weir and Cockerham's (1984) θ (henceforth referred to as F_{ST}) was calculated for each of 27 931 SNPs in the full data set between each pair of subpopulations, and across the concatenated group of all five subpopulations. SNPs within the top 2% of F_{ST} values between subpopulation pairs or across all subpopulation pairs were considered outliers, yielding a total of 11 different sets of outlier loci. Where there were multiple SNPs with identical F_{ST} values between subpopulation pairs, these SNPs were ranked secondarily by their global F_{ST} value.

Detection of directional selection on outlier loci

Based on the assumption that outlier loci would contain an excess of SNPs affected by directional selection (Lewontin and Krakauer 1973), I expected to see an excess of SNPs contained within genes (henceforth genic SNPs) in each set of outliers. To determine which SNPs were genic, the ENSEMBL Perl API was used to query the ENSEMBL database (EMBL-EBI and the Wellcome Trust Sangar Institute; http://www.ensembl.org/index.html). If a SNP was found within an intron or an exon of an annotated human gene it was classified as genic, with all remaining SNPs classified as non-genic. Because few dog genes have been annotated to date, assessing the genic status of these SNPs from homologous

human genes should provide a better overall estimate of the proportion of genic SNPs.

To determine if there was an enrichment of genic SNPs in each set of outlier loci, I performed a one-sided conditional exact test (Agresti 2002) in order to control for the ascertainment of each SNP. This test is similar to Fisher's exact test, but instead of comparing across a two-dimensional matrix, this test adjusts for a third dimension of the matrix. In this instance, I tested to see if there was a difference in the proportion of genic SNPs between the complete data set and each set of outlier loci, within each ascertainment category (grouped by species: "dog", "wolf", and "other"). All tests were performed using the standard *stats* package in R 2.9.2 (R Development Core Team 2009). Because I looked for differences across 11 different sets of outlier loci, a Bonferroni correction (Rice 1989) was applied so that P-values < 0.0045 were considered significant.

Identification of candidate genes

To identify specific loci that may have undergone selection, I examined the genomic context of the ten SNPs at highest F_{ST} between each subpopulation pair and across all subpopulations, which I will refer to as candidates of selection. Using the UCSC Genome Browser (Genome Bioinformatics group of UC Santa Cruz; http://genome.ucsc.edu/cgi-bin/hgGateway), I determined if there was any gene annotated for domestic dogs or other species at a homologous region within 1000 bases on either side of each candidate SNP, with the assumption that a SNP within 1000 bases of a gene would likely be in linkage disequilibrium (LD) with that gene. In this way, I aimed to identify genes that appeared to have undergone strong directional selection in North American grey wolves.

Results

Detection of outlier loci and directional selection

Pair-wise estimates of F_{ST} across subpopulations are tabulated in Table 3-1. These values ranged from $F_{ST} = 0.065$ between Arctic and Forest subpopulations to $F_{ST} = 0.200$ between Pacific and Baffin Island subpopulations. The global F_{ST} calculated across all subpopulations was 0.132. Eleven sets of outliers were selected as the 558 SNPs showing the highest F_{ST} between each subpopulation pair and across all subpopulations.

Across the 27 931 SNPs from the full data set, I found that 26.84% of SNPs were genic. Across all 11 sets of outlier loci, none showed a significant increase in the proportion of genic SNPs compared to the full data set (One-sided exact conditional test, 1 d.f., $P \geq 0.0087$; Table 3-2).

Identification of candidate genes

Eleven sets of candidate loci were selected as the ten SNPs showing highest F_{ST} in each pair-wise and the global F_{ST} comparison. Across all pair-wise subpopulation F_{ST} comparisons, I found that 25/100 SNPs showed up as candidate loci between >1 pair of subpopulations. Of the remaining 75 unique SNPs, I found that 24 SNPs were not within 1000 bases of an annotated gene, EST, or predicted dog gene. Ten SNPs were within 1000 bases of a predicted gene or an expressed sequence tag. This left 41 SNPs located within 1000 bases of a gene annotated in other species. Two of these SNPs were found within the same gene, and two of the 40 identified genes have been annotated in domestic dogs. These 40 genes were considered candidates of selection and are tabulated along with a brief description of any known function (recorded in OMIM and/or EntrezGene) in Table 3-3. From the global comparison, the ten SNPs with the highest F_{ST} values were found within the candidate SNPs derived from pair-wise comparisons of

subpopulations. Thus, no new candidate SNPs or genes were found by including global F_{ST} comparisons.

Discussion

Using an F_{ST}-outlier genome scan I did not find evidence for directional selection amongst outlier loci, because there was no significant enrichment for genic SNPs in any set of outliers. However, by examining the genomic context of the extreme outliers to identify genes that have potentially been historically selected, I determined that 41 out of 75 unique SNPs were found within 1000 bases of a gene annotated in another species. Following, I explore the possible reasons I did not find a signal of selection in the outlier loci, and the implications for selection on three particular genes identified near candidate SNPs.

Detection of directional selection on outlier loci

Recently, vonHoldt *et al.* (2010) used the F_{ST}-outlier approach employed here to look for genes that were important in the domestication of dogs from grey wolves. By comparing allele frequencies of ~44 000 SNPs (genotyped using the same microarray) between dogs and grey wolves, vonHoldt *et al.* (2010) found a higher proportion of genic SNPs in the top 5% of F_{ST} values. Because I used only the SNPs with the highest 2% of F_{ST} values, I expected to see high enrichment for genic SNPs in these outlier loci. However, despite looking at sets of more extreme outliers, I did not find an excess of genic SNPs, suggesting that the F_{ST} outliers did not contain a significantly higher proportion of selected loci than the entire data set. I believe the lack of signal observed can be explained by various aspects of the data set and limitations of empirical genome scans that resulted in low power to detect selection in this data set which are discussed below.

Although easy to apply, empirical genome scans for detecting selection often have a high type I error rate (Teshima *et al.* 2006) that can be exacerbated

by several factors. First, based on simulated sequence data, Teshima *et al.* (2006) found that empirical scans have a high false-discovery rate (i.e. non-selected alleles appear as outliers) when selection acts on standing (i.e. previously neutral) variation, compared to when a new allelic variant is selected. Because most SNPs interrogated on the microarray were ascertained from domestic dogs (~98% of the SNPs considered here), this is likely to be a concern. North American grey wolves are not considered to be the progenitor of the domestic dog (Savolainen *et al.* 2002; vonHoldt *et al.* 2010), and fossil evidence indicates that grey wolves have been present in North America since the Illinoian Stage of the Pleistocene (Kurtèn and Anderson 1980), which occurred between 130 000 – 300 000 years before present. This suggests that the genetic variation in dog-ascertained SNPs that are polymorphic within North American grey wolves is ≥ 130 000 years old. Thus, relative to the time-scale of selection between these subpopulations (< 12 000 years, after the receding of glaciers during the Holocene) these SNPs likely represent standing variation. It is believed that artificial selection usually acts on standing variation (Innan and Kim 2004), but it is unknown to what extent natural selection acts on new mutations compared to standing variation (Innan and Kim 2004; Hermisson and Pennings 2005): therefore it may be the case that much of the selection that has occurred between these wolf subpopulations is not represented in this data set.

Although these SNPs themselves represent ancient variation, it is possible that some are in LD with new mutations. In that case, these SNPs could also represent recent variation, which would not increase the false-discovery rate. However, it is not obvious how often such linkage is likely to occur in this data set. Assuming that wolves have a genome approximately the same size as that of the domestic dog (~2.45 Gb; Lindblad-Toh *et al.* 2005), this data set of 27 931 SNPs has a coverage of one SNP for every ~88 000 bases in the grey wolf genome. Because LD extends only a short distance in wolves (< 10 kb; Gray *et al.* 2009), much of the variation in the grey wolf genome is not detectable with this data set. Thus, the ascertainment scheme of these loci may have reduced the

power to detect selection in this data set, but the extent to which this has influenced the results is unclear.

Second, Teshima *et al.* (2006) discovered that population bottlenecks can also lead to an increase in the false-discovery rate. This may be important because Leonard *et al.* (2005) determined that the North American wolf population has undergone a dramatic reduction in population size within the past 200 years, resulting in a significant loss of genetic diversity. This suggests that the wolves in this study may have been affected by a recent bottleneck, which would therefore increase the amount of false positives detected in the genome scan.

Another limiting factor in this study is the size of my subpopulations, since three of the five subpopulations contained fewer than ten individuals. While allele frequencies of markers with few alleles are less likely to be biased by estimation from a small number of individuals (Ruzzante 1998), there is still potential for inaccurate allele frequencies with so few individuals. This would lead to inaccurate F_{ST} values, further increasing the number of false positives in the outliers. This is supported by the relatively high proportion of genic SNPs observed in the outlier SNPs between Forest and Atlantic subpopulations (30.82%), which were the only subpopulations with more than ten individuals. While not statistically greater than the proportion of genic SNPs in the complete data set (26.84%), it was the highest observed proportion of genic SNPs across all subpopulation comparisons, and was the only comparison that would have been considered significant without Bonferroni correction (One-sided conditional exact test, 1 d.f., P = 0.021). Thus, it appears that the lack of power observed is at least partially caused by small sample size.

Combined, the above considerations suggest that statistical concerns have limited my ability to detect selection in this data set. This result emphasizes the difficulty of assessing selection in wild populations even when using a large set of markers. In particular, if population demographics are unknown or if the available markers were not ascertained in the species of study, false positives in the data may obscure a signal of selection, as appears to have happened in this instance.

Further exploration of selection in this population may get better results by using other types of genome scans. In particular, scans based on haplotype structure, such as the Cross Population Extended Haplotype Homozygosity method (Sabeti *et al.* 2007; successfully employed by vonHoldt *et al.* 2010) may be more successful at detecting outliers that show evidence of directional selection, because inferences based on haplotypes rather than individual SNPs are less affected by ascertainment bias (Lohmueller *et al.* 2009).

Identification of candidate genes

Since I did not observe evidence for directional selection across SNPs in the tail end of the F_{ST} distribution within this population, it is likely that many of these loci exhibit high allele frequency differences due to genetic drift. This does not imply, however, that none of the SNPs within the extreme end of this distribution have been selected. In fact, of the 40 genes found near candidate SNPs, three have been associated with major phenotypic changes in other mammal species (*ADCY8*, *ASIP*, and *DYM*), all of which show very high allele frequency differentiation between Arctic and Forest subpopulations (Fig. 3-1).

vonHoldt *et al.* (2010) found several SNPs surrounding *ADCY8* (adenylate cyclase 8; OMIM accession 103070) that showed evidence for directional selection between dogs and wolves. This gene is implicated in memory formation in humans (de Quervain and Papassotiropoulos 2006) and behavioural sensitization in mice (Wei *et al.* 2002), leading vonHoldt *et al.* (2010) to speculate that this gene may have played an important roll in early dog domestication. Behaviour, prey type, and genetic structure have been linked in wolves [e.g. wolves that prey upon barren-ground caribou do not maintain home ranges (Musiani *et al.* 2007)] and it is possible that this gene, which appears to affect memory formation, is somehow related to the different prey types available to Forest and Arctic subpopulations (see Discussion in Chapter 2). *ADCY8* does not however appear to be related to prey specialization on the migratory caribou of

the mainland tundra because wolves on the mainland tundra (not included in the F_{ST} analyses due to high admixture) show a mix of genotypes seen in both Arctic and Forest subpopulations at this locus (data not shown). Any possible selective advantage conferred by a possible functional variant of this gene awaits further research.

DYM (*Dymeclin*; OMIM accession 607461) is a gene that is involved with bone and cartilage development. Mutations in this gene have recently been discovered to cause two developmental disorders leading to dwarfism in humans (El Ghouzzi *et al.* 2003; Neumann *et al.* 2006). In addition to being a candidate for selection between Forest and Arctic subpopulations, the intronic *DYM* SNP was found as an outlier between Arctic and Baffin subpopulations, Arctic and Atlantic subpopulations, and all subpopulations combined. This suggests that there may be a selective pressure on this gene in the Arctic subpopulation that is not experienced elsewhere. Supporting this is a study on grey wolf cranial morphology by Mulders (1997), who found that wolves exhibit a gradient of increasing skull size starting in the high arctic going southwest into the boreal forest. Notably, skull size was highly correlated with mean prey mass (Mulders 1997). Wolf size has previously been correlated with prey size (Schmitz and Lavigne 1987), and thus it is possible that the smaller available prey in the arctic selects for smaller wolves. Unfortunately there are no studies of overall wolf size across the boreal forest and arctic tundra, but the observed difference in skull sizes suggests there may be a difference in overall skeletal size between these subpopulations. While this is only a speculative correlation, it is possible that a functional difference in the *DYM* gene is related to a difference in skeletal size in these subpopulations.

More so than for the preceding two genes, there is corroborating evidence that the *ASIP* gene may have been under selection within this wolf population. Mutations in *ASIP* have been found to strongly affect light versus dark coat colouration and/or patterning in many mammal species including deer mice (*Peromyscus maniculatus*; Linnen *et al.* 2009), Soay sheep (*Ovis aries*; Gratten *et*

al. 2010), red foxes (*Vulpes vulpes*; Våge *et al.* 1997) and domestic dogs (Berryere *et al.* 2005). Coat colour varies significantly across the range of North American wolves, and recent studies have shown a strong difference in coat colour frequencies between wolves that live in the boreal forest and wolves that live in the arctic tundra (Gipson *et al.* 2002; Musiani *et al.* 2007). Tundra wolves are predominantly white, but in the forest a large proportion of wolves are black or grey, and this is thought to have adaptive significance for concealment from prey (Jolicoeur 1959). This distinction in coat colour corresponds to the differences I observed in the allele frequency of the SNP found within the *ASIP* gene, which was at high F_{ST} between boreal forest wolves (Forest and Atlantic subpopulations) and wolves found in the arctic tundra (Arctic and Baffin Island subpopulations; data not shown). Additionally, this SNP showed no differentiation in allele frequencies ($F_{ST} \sim 0$) between subpopulations within both forest and tundra habitats (data not shown). Thus, it is possible that the candidate SNP detected in this study is in linkage disequilibrium with or contained within different functional alleles of the *ASIP* gene.

I did not, however, find an association between individual coat colour and genotypes at this SNP in my sample of wolves. A subset of 29 of the wolves I genotyped were also genotyped by Musiani *et al.* (2007), and had coat colour data; however, I did not observe a correlation between phenotype [either "dark" or "light" as determined by Musiani *et al.* (2007)] and the genotype at this locus (Two-sided Fisher's exact test, 1 d.f., P = 0.830). However, this result may be biased by several factors. First, there may be a lack of signal due to noise in the colour classification system: wolf coats often vary in colour over the entire body, and Musiani *et al.* (2007) assessed colour from only a small piece of hide. Second, the subset of wolves with coat colour data was small, and was not a random subset of all wolves studied. Most of the phenotyped wolves came from the boreal forest or the mainland tundra and had either one or two copies of the major allele for this locus; there were only two phenotyped wolves that were homozygous for the minor allele. This lack of homozygotes could have a particularly strong influence on the results if a functional variant was expressed recessively.

Another consideration is that a novel allele of another gene (the *K* locus) causing dominant black coat colouration has recently made its way into the North American wolf population from domestic dogs (Anderson *et al.* 2009). In dogs, this allele (K^B) produces a protein that prevents the *Agouti* protein coded by *ASIP* from binding to its target receptor (Candille *et al.* 2007). Thus, even if *ASIP* contains functional variants within this population, the phenotypic effect may be masked in some wolves by the presence of this allele.

A final possibility is that this SNP is no longer linked to functional variation in the *ASIP* gene. A similar pattern observed in thinhorn sheep (*Ovis dalli*) was reported by Loehr *et al.* (2008), who found a strong correlation between variation within the *MC1R* coat colour gene and pelage colour at the population level, but not at the individual level. Loehr *et al.* (2008) suggested that this could be caused by incomplete genetic mixing of unlinked loci across populations or by recombination between the observed genetic variation and (separate) previously linked functional genetic variation. However, given the low genetic differentiation and high gene flow likely to occur between Arctic and Forest subpopulations (see Chapter 2), both of these explanations are unlikely.

Overall, further study is required before a strong conclusion about the selective importance of any of these three genes can be made. Because I did not detect an overall signal of selection in outlier loci, it is possible that these genes have not actually undergone selection, and their occurrence among the set of candidate genes is coincidental. Thus, it is important to emphasize that these genes remain only candidates of selection. However, because allelic variants of each of these genes have been observed to strongly affect phenotypes in other species (see above), these genes may be worth pursuing in future research to determine their involvement in North American grey wolf subpopulation differentiation.

Conclusions

I looked for a signal of directional selection within the North American wolf population, but did not find evidence for a concentration of selected markers amongst the F_{ST} outliers. Despite having a large set of markers, a number of aspects of this data set have reduced the power to detect a large-scale signal of selection in these individuals with an F_{ST}-outlier genome scan. Simply having a large marker set is not sufficient to detect selected markers, and confirms that careful study design is required for sufficient statistical power to accurately detect selection (Teshima *et al.* 2006). Further, the use of markers ascertained in model or domestic species may be of only limited use for F_{ST}-outlier scans of closely related wild species.

Despite the lack of signal and high false positive rate observed, I identified three SNPs in genes known to strongly influence phenotypes in other species that showed a high disparity in allele frequencies across Forest and Arctic subpopulations. All of these genes potentially relate to differing ecological conditions: variation in available prey type may explain selective pressure on *ADCY8* and *DYM*, while the need to blend in with a tundra compared to a forest habitat can explain selective pressure on *ASIP*. While promising, the detection of these genes is just the first step in outlining their importance to grey wolf differentiation. Re-sequencing efforts and collection of new phenotypes from across Arctic and Forest subpopulations should help to determine definitively any role these genes have in physiological adaptation to local ecological conditions.

Table 3-1. Mean pair-wise F_{ST} comparisons between five subpopulations of wolves, calculated across 27 931 SNPs.

subpopulation	Arctic	Atlantic	Baffin	Forest	Pacific
Arctic	0.000				
Atlantic	0.110	0.000			
Baffin	0.116	0.145	0.000		
Forest	0.065	0.073	0.103	0.000	
Pacific	0.168	0.166	0.200	0.111	0.000

Table 3-2. Proportion SNPs in each group of outlier loci (n = 558) that were found within homologs of annotated human genes. No set of outlier loci was found to contain a significantly greater proportion of genic SNPs compared to the complete set of 27 931 SNPs after Bonferroni correction (One-sided exact conditional test, 1 d.f., $P \geq 0.0087$).

	% of SNPs found within genes	**P-value**
All 27 931 SNPs	**26.84**	
Subpopulation comparisons containing F_{ST} outliers		
Atlantic/Arctic	27.06	0.471
Baffin/Arctic	25.99	0.689
Baffin/Atlantic	25.99	0.689
Pacific/Arctic	23.11	0.979
Pacific/Atlantic	28.49	0.206
Pacific/Baffin	24.37	0.912
Forest/Arctic	24.91	0.853
Forest/Atlantic	30.82	0.021
Forest/Baffin	27.06	0.467
Forest/Pacific	28.49	0.205
All subpopulations	**28.49**	**0.207**

Table 3-3. Name, location, function, and putative role of 40 genes found near candidate loci. Location is that of the identified SNP.

Gene	Chromosome	Location	Function	Putative role/location
ACSL3	37	31898472	lipid biosynthesis, degradation	highly expressed in brain
ADCY8	13	30806609	catalyses formation of cAMP from ATP	involved with memory formation
ARMC9	25	46157665	binding	uncertain
ASIP	24	26359293	binds to melanocortin receptor	affects pelage colouration
ASTN2	11	73462161	protein binding	involved in neuronal migration
BCAS1	24	42710476	uncertain	possible oncogene
CACNA1S	27	47092881	subunit of calcium channel	expressed in skeletal muscle cells
CLEG4G	20	55481308	protein/sugar binding	uncertain
CPVL	14	45047414	carboxypeptidase	expressed by maturing monocytes
CSMD3	13	15478348	uncertain	associated with Epilepsy
DCLK1	25	7668360	microtubule-associated kinase	uncertain
DYM	7	82484918	uncertain	mutations cause dwarfism/retardation
EEPD1	14	50645084	DNA binding	DNA repair
Fam170a	11	11900186	metal ion binding	uncertain
FAM65B	35	25908532	binding	uncertain
FOXN3	8	63593443	transcription factor	uncertain
GRIA1	4	59029131	glutamate receptor	involved with neurotransmission
ICOS	37	15816476	protein binding	involved in T-cell receptor signaling
LRRC16A	35	26355163	inhibits actin filament capping	uncertain
MFC2L	22	63476100	activates GTP-binding proteins	uncertain

Gene	Chromosome	Location	Function	Putative role/location
NPAS3	8	15646515, 15785878	transcription factor	involved with neurogenesis, schizophrenia
NRG3	4	34574181	tyrosine phosphorylation	associated with schizophrenia
PAPPA	11	73113432	cleaves IGF binding proteins	associated with wound healing
PCNXL3	4	9730078	uncertain	uncertain
PIK3R5	5	36551475	subunit of kinase regulator	uncertain
PLCL2	23	29027820	calcium ion binding	cell signaling
PREP	12	65760166	cleaves peptide hormones	involved in mediating sperm death
PTPN14	7	15319772	phosphatase	involved with cytoskeleton
RPS27L	18	33189957	interacts with apoptosis gene	regulates apoptosis for cancer suppression
RXRB	9	53860854	retinoid receptor	involved with spermogenesis
slc38a2	27	10944564	amino acid transporter	uncertain
SNAPC4	9	52439042	DNA binding	snRNA transcription
SNX29	6	33613786	protein binding	cell communication
TBCE	4	7525804	chaperones microtubule folding	associated with retardation, osteosclerosis
TIGD3	18	54968633	transposon	uncertain
TMEM132C	26	6092231	uncertain	uncertain
TRIB2	17	12221636	signal transduction	expressed in thyroid
USH2A	38	14317121	uncertain	involved in development of inner ear
ZSCAN12	35	28545384	regulation of transcription	expressed in testes

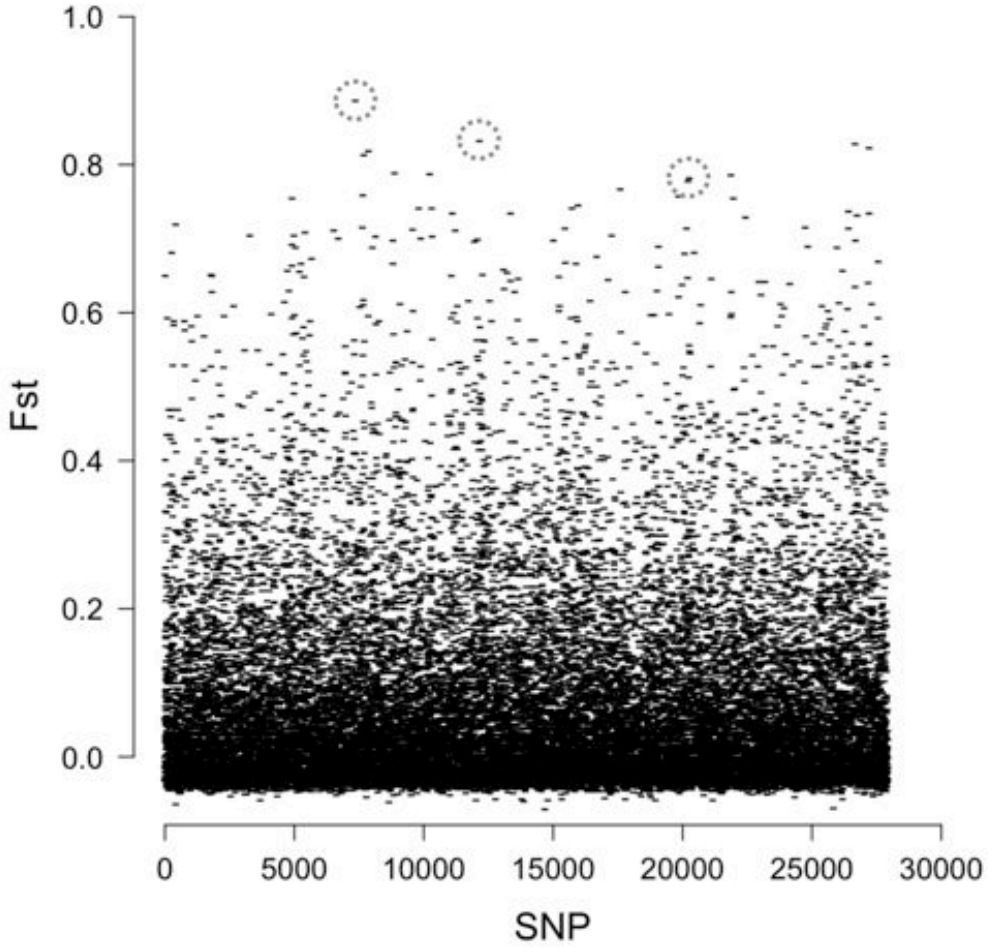


Figure 3-1. F_{ST} values of 27 931 SNPs between Arctic and Forest subpopulations, showing high differentiation of SNPs located in the *DYM* gene (circled, left), the *ADCY8* gene (centre) and the *ASIP* gene (right). Spacing of SNPs along the x-axis is not to scale of the distance between SNPs along chromosomes; SNPs are ordered primarily by chromosome number and secondarily by position within the chromosome.

Chapter 4
Synthesis

Grey wolves, being a top predator throughout most of their range, are highly important members of northern ecosystems. They can disperse very long distances (Fritts 1983; Wabakken *et al.* 2007), but despite this, grey wolf populations appear to be highly structured at large scales, even in the absence of physical barriers (Pilot *et al.* 2006; Carmichael *et al.* 2007). Interestingly, there has been no previous evidence for strong isolation by distance at a continental scale in Europe (Pilot *et al.* 2006) or North America (Carmichael *et al.* 2007), nor has genetic differentiation been strongly correlated with distance at finer scales (Weckworth *et al.* 2005; Musiani *et al.* 2007; Muñoz-Fuentes *et al.* 2009). Instead, the structure of grey wolf populations is organized based on local ecological factors including habitat, climate, and available prey (Weckworth *et al.* 2005; Pilot *et al.* 2006; Muñoz-Fuentes *et al.* 2009). In particular, Carmichael *et al.* (2001; 2007) and Musiani *et al.* (2007) have found the presence of migratory prey to be particularly influential on population structure.

Most previous studies of grey wolf population structure to date have been performed using microsatellite markers. While these markers are relatively easy to obtain, inexpensive, and exhibit high allelic polymorphism, they generally have low overall genomic coverage. SNP markers, which have much higher genomic coverage and are less prone to sampling bias due to low allelic polymorphism (Ruzzante *et al.* 1998), should prove to be useful in studies of population genetics and genomics (Morin *et al.* 2004). Additionally, due to the high genomic coverage, SNP markers can be used to detect selection and identify genes that have been important in phenotypic adaptation (Morin *et al.* 2004). While most large-scale SNP studies have involved humans (e.g. Novembre *et al.* 2008; Pickrell *et al.* 2009), large SNP panels have also been developed for other species including sheep, cattle, and domestic dogs. The use of a domestic dog SNP array

to genotype grey wolves was verified by vonHoldt *et al.* (2010), but there has been no previous effort to look at genetic structure in a wild population using a large SNP data set.

Using a large SNP panel developed from domestic dogs, I genotyped grey wolves from across Canada and Alaska, accomplishing four goals: 1) Re-evaluate population structure of North American grey wolves using a high-density marker. 2) Compare SNP results to microsatellite-based results, to determine possible advantages of SNP data. 3) Explore how many SNPs are necessary to accurately assess population structure. 4) Scan the genome for signs of selection to identify genes that may have adaptive consequences for wolves in different habitats.

Conclusions

In Chapter 2, I explored the first three goals. I identified only five genetic subpopulations of wolves across North America, compared to the eight subpopulations suggested by Carmichael *et al.* (2007). However, four of the subpopulations I detected correspond to subpopulations identified by Carmichael *et al.* (2007), and support the ecological separation reported therein. Notably, a high number of wolves (45/132) were not strongly assigned to any single subpopulation, indicating that a significant amount of gene flow occurs between subpopulations, which was not inferred from microsatellite data (Carmichael *et al.* 2007). In particular, almost all the individuals captured on the mainland tundra (north of the tree line, Fig. 2-2a) were highly admixed between Arctic and Forest subpopulations. Wolves in this area have previously been identified as a distinct subpopulation (Carmichael *et al.* 2007; Musiani *et al.* 2007), and it was hypothesized that specialization on migratory barren-ground caribou in this region was the driving force behind the observed population structure. My result indicates that rather than differentiating these wolves, specialization on migratory prey is causing increased gene flow across the mainland tundra, reducing differentiation between Arctic and Forest subpopulations.

I found that the genetic structure of North American wolves only partially corresponds to the subspecies distribution suggested by Nowak (1995). This indicates that skull morphology may not always be an accurate indicator of subspecies. Minimally, it confirms Carmichael *et al.*'s (2007) finding that the structure of North American grey wolves is more highly influenced by contemporary gene flow than by the previous separation of subspecies into different glacial refugia.

Contrary to previous results (Pilot *et al.* 2006; Carmichael *et al.* 2007), I found strong evidence that geographic distance influences grey wolf genetic structure, which helps to explain the reduced structure [compared to Carmichael *et al.* (2007)] and high number of admixed individuals observed. By comparing pair-wise inter-individual genetic distance estimates, I found a much stronger signal of isolation by distance in SNP data than microsatellite data in a common set of 61 previously-genotyped wolves. This indicates that microsatellites lack power to detect isolation by distance, which is likely a result of noise due to the high allelic polymorphism of these markers. I also demonstrated that this SNP data set has high resolution, by recovering substructure within a subpopulation containing only eight individuals, and finding evidence for isolation by distance within three of five subpopulations. This confirms that large SNP panels can be used to detect population structure even when very few samples are available for genotyping, a finding that may be useful for studies of rare or endangered species.

Finally, I found that decreasing the number of markers used to infer structure yielded results that were increasingly discordant to results from the full data set. Equally important, I observed high variance in discordance across small data subsets containing the same number of SNPs. This may be a concern for studies using a limited number of markers, because there is potential for high mis-assignment when using smaller data sets. However, I found that increasing the number of SNPs had decreasing benefits as the number of SNPs approached the full data set. This suggests that genotyping ~500 SNPs may approximate a reasonable balance between accuracy of assignment and cost of genotyping. Last,

I discovered that highly admixed individuals are most likely to be mis-assigned, and that overall admixture was underestimated in small data sets. This means that studies looking to assess or quantify genetic admixture in particular should use as many markers as possible in order to minimize error in assignment of admixed individuals, and to make sure that admixture is correctly identified.

In Chapter 3, I addressed the final goal of my study, looking for genes that may be under selection in this population of wolves. By using an empirical F_{ST}-outlier genome scan, I looked for SNPs showing particularly high differentiation between subpopulation pairs, assuming that SNPs that have undergone directional selection between subpopulations would show extreme F_{ST} values between subpopulations. Looking at the 2% outliers from the pairwise and global F_{ST} distributions, I did not find evidence for an increased proportion of markers that have undergone directional selection, because these outliers did not show an increased proportion of genic SNPs compared to the complete set of SNPs. This is likely the result of low power to detect selected loci in this population due to a) the fact that these SNPs, being mostly ascertained from domestic dogs, may represent standing variation, b) the recent reduction in size of the North American wolf population, and c) the small sample sizes of several subpopulations. This result emphasizes a need to carefully select markers and individuals when performing genome scans, and suggests that loci ascertained in domestic species may be inappropriate for use in F_{ST}-outlier scans of wild species.

However, by looking at the extreme outliers, I identified 41 unique SNPs near to or contained within genes annotated in other species. Of these, three have been found to have significant phenotypic effect in other species, and were found at very high F_{ST} between Arctic and Forest subpopulations. Two of these genes, which are involved with memory (*ADCY8*; de Quervain and Papassotiropoulos 2006) and skeletal development (*DYM*; El Ghouzzi *et al.* 2003), may be differentially selected between Forest and Arctic subpopulations due to the different prey types available to each subpopulation. There is previous evidence showing a change in skull size between the boreal forest and the high Arctic

tundra (Mulders 1997), lending support to a hypothesis of selection on the *DYM* gene. The third gene (*ASIP*) has been observed to control fur colouration in many mammal species including domestic dogs (Berryere *et al.* 2005) and red foxes (Våge *et al.* 1997). Notably, the frequency difference in this SNP is correlated with a high frequency of dark-coloured wolves in the boreal forest and white wolves in Arctic tundra habitats (Gipson *et al.* 2002; Musiani *et al.* 2007). This colour change is likely locally adaptive, allowing wolves to blend with their surroundings to avoid detection by prey (Jolicoeur 1959). I must however emphasize again that these genes are only candidates of selection, and further study will be necessary before their importance to grey wolf phenotypic differentiation can be verified.

Future directions

In this study I have looked at only a small subset of the analyses available to population geneticists. To fully appreciate the benefits of using a large panel of SNPs for population genetic analysis, many more types of analysis could be performed. For this study system, it would be interesting to estimate recent and historical migration rates between subpopulations, as well as spatial autocorrelation, as both would help us to better understand the observed pattern of isolation by distance. Additionally, an assessment of recent migration rates between Arctic and Forest subpopulations could verify the conclusion of high gene flow presented here.

Although I have made useful comparisons regarding assignment based on varying numbers of markers, determining the full effect of the number of markers on subpopulation assignment requires further study. In particular, there are three important questions that remain unanswered. First, what is the relative importance of the number of markers compared to the number of sampled individuals? STRUCTURE assignments are affected by both the number of individuals and the number of markers (see Chapter 2), and it is not clear which is more important.

Second, how likely is STRUCTURE to determine the correct number of subpopulations when using data sets with few markers? While I assessed their relative accuracy with a fixed number of subpopulations, it is possible that smaller data sets would not indicate the correct number of subpopulations, thus making my estimates of discordance conservative. Third, how many SNPs are equivalent to microsatellites for assignment purposes? Unfortunately I had access to only a small number of individuals genotyped with a limited number of microsatellites. The answer to this question will require a more complete set of individuals genotyped at many more microsatellite loci, so that multiple replicates can be compared, as was done for SNPs in this study.

I did not recover a signal of natural selection in this population, which suggests that many markers near selected genes may have been missed by the empirical genome scan. This could potentially be overcome by genotyping a greater number of individuals, which would help to decrease the number of false positive markers found in the outliers, and provide stronger evidence that extreme outliers were selected. Additionally, the use of haplotype-based genome scans may help to detect regions under selection. In particular, the Cross Population Extended Haplotype Homozygosity test (Sabeti *et al.* 2007) may prove useful, since it (unlike several other LD-based genome scans) requires no estimates of linkage disequilibrium. This test was employed successfully by vonHoldt *et al.* (2010) to domestic dogs, and may also be successful at recovering genes under selection in wolves.

While I did not directly include ecological variables in any analyses, it may be useful to obtain environmental data at the capture location of each individual (such as temperature, precipitation, and vegetation type) in order to look for genetic correlates of ecological variables. In conjunction with a genome scan, methods such as the spatial analysis method (SAM; Joost *et al.* 2008) or the Bayesian geographic analysis of Hancock *et al.* (2008) could be used to identify markers associated with particular ecological variables. By combining the results of a genome scan with a correlation analysis, it should be possible to discover

SNPs that have been under selection, and precisely identify environmental variables correlated with the selective pressure on each SNP. Finding SNPs that are correlated with particular environmental variables across subpopulations will help locate genes associated with ecological selective pressures that are not detectable when comparing inter-subpopulation allele frequencies.

Last, the possible importance of *ADCY8*, *ASIP*, and *DYM* to grey wolf phenotypic differentiation have yet to be empirically assessed. For *ADCY8* in particular, such an assessment will be difficult at this time, because *ADCY8* is only one gene in a particular set found to be related to memory formation (de Quervain and Papassotiropoulos 2006). Determining the possible function of the other two genes will require sequencing these genes across numerous individuals, as well as collecting phenotypes for skeletal morphology and development time, and obtaining more precise measurements of wolf pelage colour. This is an expensive proposition, so a logical first step would be re-genotyping individuals at the detected SNPs in order to confirm genotypes. Hopefully, through a more comprehensive interrogation of the DNA sequence surrounding each SNP and careful phenotypic characterization, we will be able to determine a causative relationship for particular wolf phenotypes.

Literature Cited

Agresti A (2003) Dealing with discreteness: making exact confidence intervals for proportions, differences of proportions, and odds ratios more exact. Statistical Methods in Medical Research 12, 3-21.

Aitken N, Smith S, Schwarz C, Morin PA (2004) Single nucleotide polymorphism (SNP) discovery in mammals: a targeted-gene approach. Molecular Ecology 13, 1423-1431.

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

Allendorf FW, Leary RF, Spruell P, Wenburg J (2001) The problems with hybrids: setting conservation guidelines. Trends in Ecology & Evolution 16, 613-622.

Altschul S, Gish W, Miller W, Myers E, Lipman DJ (1990) Basic local alignment search tool. Journal of Molecular Biology 215, 403-410.

Anand-Wheeler I (2002) Terrestrial Mammals of Nunavut. Department of Sustainable Development, Nunavut Wildlife Management Board, Iqaluit, Canada.

Anderson TM, vonHoldt BM, Candille SI *et al.* (2009) Molecular and Evolutionary History of Melanism in North American Gray Wolves. Science 323, 1339-1343.

Aspi J, Roininen E, Kiiskilä J *et al.* (2009) Genetic structure of the northwestern Russian wolf populations and gene flow between Russia and Finland. Conservation Genetics 10, 815-826.

Beaumont M, Nichols R (1996) Evaluating loci for use in the genetic analysis of population structure. Proceedings of the Royal Society of London B: Biological Sciences 263, 1619-1626.

Berryere TG, Kerns JA, Barsh GS, Schmutz SM (2005) Association of an Agouti allele with fawn or sable coat color in domestic dogs. Mammalian Genome 16, 262-272.

Bishop D, Demenais F, Iles M *et al.* (2009) Genome-wide association study identifies three loci associated with melanoma risk. Nature Genetics 41, 920-925.

Black WC, Baer CF, Antolin MF, DuTeau NM (2001) Population genomics: genome-wide sampling of insect populations. Annual Review of Entomology 46, 441-469.

Boitani L (1995) Ecological and cultural diversities in the evolution of wolf-human relationships. In: Ecology and conservation of wolves in a changing world (eds. Carbyn L, Fritts S, Seip D), pp. 3-11. Canadian Circumpolar Institute, Edmonton, Canada.

Boitani L (2003) Wolf conservation and recovery. In: Wolves: Behavior, Ecology, and Conservation (eds. Mech LD, Boitani L), pp. 317-340. University of Chicago Press, Chicago.

Brandström M, Ellegren H (2008) Genome-wide analysis of microsatellite polymorphism in chicken circumventing the ascertainment bias. Genome Research 18, 881-887.

Brumfield RT, Beerli P, Nickerson DA, Edwards SV (2003) The utility of single nucleotide polymorphisms in inferences of population history. Trends in Ecology & Evolution 18, 249-256.

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

Carmichael LE, Krizan J, Nagy JA *et al.* (2008) Northwest passages: conservation genetics of Arctic Island wolves. Conservation Genetics 9, 879-892.

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

Carmichael LE, Nagy JA, Larter NC, Strobeck C (2001) Prey specialization may influence patterns of gene flow in wolves of the Canadian Northwest. Molecular Ecology 10, 2787-2798.

de Quervain DJ, Papassotiropoulos A (2006) Identification of a genetic cluster influencing memory performance and hippocampal activity in humans. Proceedings of the National Academy of Sciences 103, 4270-4274.

di Rienzo A, Peterson AC, Garza JC *et al.* (1994) Mutational processes of simple-sequence repeat loci in human populations. Proceedings of the National Academy of Sciences 91, 3166-3170.

El Ghouzzi V, Dagoneau N, Kinning E *et al.* (2003) Mutations in a novel gene Dymeclin (FLJ20071) are responsible for Dyggve-Melchior-Clausen syndrome. Human Molecular Genetics 12, 357-364.

Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. Heredity 103, 285-298.

Excoffier L, Laval G, Schneider S (2005) Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evolutionary Bioinformatics Online 1, 47-50.

Excoffier L, Slatkin M (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Molecular Biology and Evolution 12, 921-927.

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164, 1567-1587.

Faubet P, Waples RS, Gaggiotti OE (2007) Evaluating the performance of a multilocus Bayesian method for the estimation of migration rates. Molecular Ecology 16, 1149-1166.

Fritts SH (1983) Record dispersal by a wolf from Minnesota. Journal of Mammalogy 64, 166-167.

Gautier M, Flori L, Riebler A *et al.* (2009) A whole genome bayesian scan for adaptive genetic divergence in West African cattle. BMC Genomics 10, 550.

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

Gipson PS, Bangs EE, Bailey TN *et al.* (2002) Color patterns among wolves in western North America. Wildlife Society Bulletin 30, 821-830.

Gore M, Wright M, Ersoz E, Bouffard P (2009) Large-scale discovery of gene-enriched SNPs. The Plant Genome 2, 121-133.

Gratten J, Pilkington JG, Brown E *et al.* (2010) The genetic basis of recessive self-colour pattern in a wild sheep population. Heredity 104, 206-214.

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

Groves CP (1993) A theory of human and primate evolution. Clarendon Press, Oxford, England.

Guillot G, Leblois R, Coulon A, Frantz AC (2009) Statistical methods in spatial genetics. Molecular Ecology 18, 4734-4756.

Hall ER (1981) The mammals of North America. John Wiley and Sons, New York.

Hanage WP, Fraser C, Tang J, Connor TR, Corander J (2009) Hyper-recombination, diversity, and antibiotic resistance in pneumococcus. Science 324, 1454-1457.

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

Innan H, Kim Y (2004) Pattern of polymorphism after strong artificial selection in a domestication event. Proceedings of the National Academy of Sciences 101, 10667-10672.

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

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

Joost S, Kalbermatten M, Bonin A (2008) Spatial analysis method (SAM): a software tool combining molecular and environmental data to identify candidate loci for selection. Molecular Ecology Resources 8, 957-960.

Kijas JW, Townley D, Dalrymple BP *et al.* (2009) A genome wide survey of SNP variation reveals the genetic structure of sheep breeds. PLoS One 4, e4668.

Kimura M (1969) The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics 61, 893-903.

Kimura M, Ohta T (1978) Stepwise mutation model and distribution of allelic frequencies in a finite population. Proceedings of the National Academy of Sciences 75, 2868-2872.

Kurtén B, Anderson E (1980) Pleistocene mammals of North America. Columbia University Press, New York.

Lao O, Lu TT, Nothnagel M *et al.* (2008) Correlation between genetic and geographic structure in Europe. Current Biology 18, 1241-1248.

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

Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics 74, 175-195.

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

Linnen CR, Kingsley EP, Jensen JD, Hoekstra HE (2009) On the origin and spread of an adaptive allele in deer mice. Science 325, 1095-1098.

Loehr J, Worley K, Moe J, Carey J, Coltman DW (2008) MC1R variants correlate with thinhorn sheep colour cline but not individual colour. Canadian Journal of Zoology 86, 147-150.

Lohmueller KE, Bustamante CD, Clark AG (2009) Methods for human demographic inference using haplotype patterns from genomewide single-nucleotide polymorphism data. Genetics 182, 217-231.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. Nature Reviews Genetics 4, 981-994.

Mantel N (1967) The detection of disease clustering and a generalized regression approach. Cancer research 27, 209-220.

Mariette S, Le Corre V, Austerlitz F, Kremer A (2002) Sampling within the genome for measuring within-population diversity: trade-offs between markers. Molecular Ecology 11, 1145-1156.

Marth GT, Korf I, Yandell MD *et al.* (1999) A general approach to single-nucleotide polymorphism discovery. Nature Genetics 23, 452-456.

Mech, LD (1970) The Wolf: The ecology and behavior of an endangered species. Natural History Press, Doubleday Publishing Company, New York.

Mech LD (2005) Decline and recovery of a High Arctic wolf-prey system. Arctic 58, 305-307.

Mech LD, Adams L, Meier T, Burch J, Dale B (1998) The wolves of Denali. University of Minnesota Press, Minneapolis.

Mech LD, Boitani L (2003) Wolf social ecology. In: Wolves: Behavior, Ecology, and Conservation (eds. Mech LD, Boitani L), pp. 1-34. University of Chicago Press, Chicago.

Mech LD, Boitani L (2004a) Chapter 4.1: Coyote. In: Canids: foxes, wolves, jackals and dogs. Status survey and conservation action plan (eds. Sillero-Zubiri C, Hoffman M, MacDonald D), pp. 81-87. IUCN Canid Specialist Group, Gland, Switzerland.

Mech LD, Boitani L (2004b) Chapter 5.3: Grey Wolf. In: Canids: foxes, wolves, jackals and dogs. Status survey and conservation action plan (eds. Sillero-Zubiri C, Hoffman M, MacDonald D), pp. 124-129. IUCN Canid Specialist Group, Gland, Switzerland.

Morin PA, Luikart G, Wayne RK, the SNP workshop group (2004) SNPs in ecology, evolution and conservation. Trends in Ecology & Evolution 19, 208-216.

Mulders R (1997) Geographic variation in the cranial morphology of the wolf (*Canis lupus*) in northern Canada. M.Sc. Thesis, Laurentian University, Sudbury, Canada.

Muñoz-Fuentes V, Darimont CT, Wayne RK, Paquet PC, Leonard JA (2009) Ecological factors drive differentiation in wolves from British Columbia. Journal of Biogeography 36, 1516-1531.

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

Neff BD, Gross MR (2001) Microsatellite evolution in vertebrates: inference from AC dinucleotide repeats. Evolution 55, 1717-1733.

Nei M (1972) Genetic distance between populations. The American Naturalist 106, 283-292.

Neumann LM, El Ghouzzi V, Paupe V *et al.* (2006) Dyggve-Melchior-Clausen syndrome and Smith-McCort dysplasia: Clinical and molecular findings in three families supporting genetic heterogeneity in Smith-McCort dysplasia. American Journal of Medical Genetics 140A, 421-426.

Nielsen R (2005) Molecular signatures of natural selection. Annual Review of Genetics 39, 197-218.

Nielsen R, Hubisz MJ, Clark AG (2004) Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. Genetics 168, 2373-2382.

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

Novembre J, Johnson T, Bryc K *et al.* (2008) Genes mirror geography within Europe. Nature 456, 98-101.

Nowak RM (1979) North American quaternary *Canis*. Monograph no. 6. Museum of Natural History, University of Kansas, Lawrence.

Nowak RM (1995) Another look at wolf taxonomy. In: Ecology and conservation of wolves in a changing world (eds. Carbyn L, Fritts S, Seip D), pp. 375-398. Canadian Circumpolar Institute, Edmonton, Canada.

Nowak RM (2003) Wolf evolution and taxonomy. In: Wolves: Behavior, Ecology, and Conservation (eds. Mech LD, Boitani L), pp. 239–258. University of Chicago Press, Chicago.

Nübel U, Dordel J, Kurt K *et al.* (2010) A timescale for evolution, population expansion, and spatial spread of an emerging clone of methicillin-resistant *Staphylococcus aureus*. PLoS Pathogens 6, e1000855.

Ochieng JW, Shepherd M, Baverstock PR *et al.* (2008) Genetic variation within two sympatric spotted gum eucalypts exceeds between taxa variation. Silvae Genetica 57, 249-256.

Oksanen J, Kindt R, Legendre P *et al.* (2009) vegan: Community Ecology Package. R package version 1.15-4. http://CRAN.R-project.org/package=vegan.

Pant SD, Schenkel FS, Verschoor CP *et al.* (2010) A principal component regression based genome wide analysis approach reveals the presence of a novel QTL on BTA7 for MAP resistance in holstein cattle. Genomics 95, 176-182.

Park SDE (2001) Trypanotolerance in West African cattle and the population genetic effects of selection. Ph.D. thesis, University of Dublin, Dublin, Ireland.

Parker G (1973) Distribution and densities of wolves within barren-ground caribou range in northern mainland Canada. Journal of Mammalogy 54, 341-348.

Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genetics 2, e190.

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

Pertoldi C, Tokarska M, Wojcik JM *et al.* (2009) Depauperate genetic variability detected in the American and European bison using genomic techniques. Biology Direct 4, 48.

Pickrell JK, Coop G, Novembre J *et al.* (2009) Signals of recent positive selection in a worldwide sample of human populations. Genome Research 19, 826-837.

Pilot M, Jedrzejewski W, Branicki W *et al.* (2006) Ecological factors influence population genetic structure of European grey wolves. Molecular Ecology 15, 4533-4553.

Pompanon F, Bonin A, Bellemain E, Taberlet P (2005) Genotyping errors: causes, consequences and solutions. Nature Reviews Genetics 6, 847-859.

Price AL, Patterson NJ, Plenge RM *et al.* (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics 8, 904-909.

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

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

R Development Core Team (2009) R: A language and environment for statistical computing. R Foundation for StatisticalComputing, Vienna, Austria. http://www.R-project.org.

Rice WR (1989) Analyzing tables of statistical tests. Evolution 43, 223-225.

Roden SE, Dutton PH, Morin PA (2009) AFLP fragment isolation technique as a method to produce random sequences for single nucleotide polymorphism discovery in the green turtle, *Chelonia mydas*. Journal of Heredity 100, 390-393.

Rokas A, Williams BL, King N, Carroll SB (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425, 798-804.

Rosenberg NA, Mahajan S, Ramachandran S *et al.* (2005) Clines, clusters, and the effect of study design on the inference of human structure. PLoS Genetics 1, e70.

Roy MS, Geffen E, Smith D, Ostrander EA, Wayne RK (1994) Patterns of differentiation and hybridization in North American wolflike canids, revealed by analysis of microsatellite loci. Molecular Biology and Evolution 11, 553-570.

Ruzzante DE (1998) A comparison of several measures of genetic distance and population structure with microsatellite data: bias and sampling variance. Canadian Journal of Fisheries and Aquatic Sciences 55, 1-14.

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

Santure AW, Stapley J, Ball AD *et al.* (2010) On the use of large marker panels to estimate inbreeding and relatedness: empirical and simulation studies of a pedigreed zebra finch population typed at 771 SNPs. Molecular Ecology 19, 1439-1451.

Savolainen P, Zhang Y, Luo J, Lundeberg J, Leitner T (2002) Genetic evidence for an East Asian origin of domestic dogs. Science 298, 1610-1613.

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

Schmitz OJ, Lavigne DM (1987) Factors affecting body size in sympatric Ontario *Canis*. Journal of Mammalogy 68, 92-99.

Schwartz MK, McKelvey KS (2009) Why sampling scheme matters: the effect of sampling scheme on landscape genetic results. Conservation Genetics 10, 441-452.

Silva-Zolezzi I, Hidalgo-Miranda A, Estrada-Gil J *et al.* (2009) Analysis of genomic diversity in Mexican Mestizo populations to develop genomic medicine in Mexico. Proceedings of the National Academy of Sciences 106, 8611-8616.

Simonsen KL, Churchill GA, Aquadro CF (1995) Properties of statistical tests of neutrality for DNA polymorphism data. Genetics 141, 413-429.

Slate J (2005) Quantitative trait locus mapping in natural populations: progress, caveats and future directions. Molecular Ecology 14, 363-379.

Slate J, Gratten J, Beraldi D *et al.* (2009) Gene mapping in the wild with SNPs: guidelines and future directions. Genetica 136, 97-107.

Smouse PE, Long JC, Sokal RR (1986) Multiple regression and correlation extenstions of the Mantel test of matrix correspondence. Systematic Biology 35, 627-632.

Stinchcombe JR, Hoekstra HE (2007) Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. Heredity 100, 158-170.

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

Tang J, Vosman B, Voorrips RE, Linden CGvd, Leunissen JAM (2006) QualitySNP: a pipeline for detecting single nucleotide polymorphisms and insertions/deletions in EST data from diploid and polyploid species. BMC Bioinformatics 7, 438.

Teshima KM, Coop G, Przeworski M (2006) How reliable are empirical genomic scans for selective sweeps? Genome Research 16, 702-712.

Våge DI, Lu D, Klungland H *et al.* (1997) A non-epistatic interaction of agouti and extension in the fox, Vulpes vulpes. Nature Genetics 15, 311-315.

Väli Ü, Einarsson A, Waits L, Ellegren H (2008) To what extent do microsatellite markers reflect genome-wide genetic diversity in natural populations? Molecular Ecology 17, 3808-3817.

Venter JC, Adams MD, Meyers EW *et al.* (2001) The sequence of the human genome. Science 291, 1304-1351.

Vignal A, Milan D, SanCristobal M, Eggen A (2002) A review on SNP and other types of molecular markers and their use in animal genetics. Genetics Selection Evolution 34, 275-305.

Vilà C, Savolainen P, Maldonado JE *et al.* (1997) Multiple and ancient origins of the domestic dog. Science 276, 1687-1689.

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

Voight B, Kudaravalli S, Wen X, Pritchard J (2006) A map of recent positive selection in the human genome. PLoS Biology 4, 446-458.

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

Wabakken P, Sand H, Kojola I *et al.* (2007) Multistage, long-range Natal dispersal by a global positioning system–collared Scandinavian wolf. Journal of Wildlife Management 71, 1631-1634.

Walton LR, Cluff HD, Paquet PC, Ramsay MA (2001) Movement patterns of barren-ground wolves in the central Canadian Arctic. Journal of Mammalogy 82, 867-876.

Wang ET, Kodama G, Baldi P, Moyzis RK (2006) Global landscape of recent inferred Darwinian selection for Homo sapiens. Proceedings of the National Academy of Sciences 103, 135-140.

Wang X, Tedford RH (1994) Basicranial anatomy and phylogeny of primitive canids and closely related miacids (Carnivora: Mammalia). American Museum Novitiates #3092, American Museum of Natural History, New York.

Wayne RK, Van Valkenburgh B, O'Brien SJ (1991) Molecular distance and divergence time in carnivores and primates. Molecular Biology and Evolution 8, 297-319.

Weckworth BV, Talbot S, Sage GK, Person DK, Cook J (2005) A signal for independent coastal and continental histories among North American wolves. Molecular Ecology 14, 917-931.

Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. Evolution 38, 1358-1370.

Wozencraft WC (1993) Order Carnivora. In: Mammal species of the world: A taxonomic and geographic reference (eds. Wilson DE, Reeder DE), pp. 279-348. Smithsonian Institution Press, Washington, D.C.

Wright S (1951) The genetical structure of populations. Annals of Eugenics 15, 323-354.