DISSERTATION

MODELING GENETIC CORRELATION IN MICROSATELLITE FREQUENCIES ASSOCIATED WITH COVARIATES AND POPULATION SUBSTRUCTURE

Submitted by

Isin Ozaksoy

Statistics Department

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Spring, 2007

COLORADO STATE UNIVERSITY

December 18, 2006

WE HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER OUR SUPERVISION BY ISIN OZAKSOY ENTITLED MODELING GENETIC CORRELATION IN MICROSATELLITE FREQUENCIES ASSOCIATED WITH COVARIATES AND POPULATION SUBSTRUCTURE BE ACCEPTED AS FULFILLING IN PART REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY.

Committee on Graduate Work

Advisor

Department Head

ABSTRACT OF DISSERTATION

MODELING GENETIC CORRELATION IN MICROSATELLITE FREQUENCIES ASSOCIATED WITH COVARIATES AND POPULATION SUBSTRUCTURE

The survival of an endangered species can depend on how accurately the population structure of that species is identified. By determining the substructuring of a species, wise management can be facilitated. A popular tool for the detection and estimation of population structure is information extracted from genotypic data from individuals within populations.

In this dissertation I develop a new method that models genetic correlation structure and relates it to a covariate. Two sources of structure are isolated: (1) substructuring of a population into genetically distinct substocks, and (2) genetic correlation within a substock corresponding to a measurable covariate. My modeling approach is based on match probabilities for different types of allele pairs, and on marginalization of a beta-binomial probability model. My statistical model can be fit by adapting GAM fitting methodologies. Hypotheses can be tested using permutation methods.

In order to evaluate the performance of my method I examine diverse simulations. I consider both one-population and two-population cases. In the one-population case, within-stock correlation attributable to a covariate is the influential factor, while for the two-population case both within-substock correlation and population substructuring can be detected. In these studies, I analyze the influence of various simulation parameters and compare the performance of my method with other related methods. Generally, my method

is shown to have good power to detect all but the tiniest effect sizes in datasets limited to a small number of loci and samples.

I also examine the performance of my method by applying it to real data. The two examples I consider pertain to the Bering-Chukchi-Beaufort Seas stock of bowhead whales and to black-tailed prairie dogs living in northern Colorado. In both cases application of my method is found to corroborate results from previous research.

The dissertation concludes with a discussion of some of the strengths and weaknesses of my approach, and some consideration of potential future research.

Isin Ozaksoy

Statistics Department

Colorado State University

Fort Collins, CO 80523

Spring 2007

ACKNOWLEDGEMENTS

Above all, I would like to thank my advisor, Dr. Geof H. Givens for his support and guidance. I would also like to thank my committee members, Dr. Micheal F. Antolin, Dr. F. Jay Breidt and Dr. Brad J. Biggerstaff.

I would like to acknowledge the North Slope Borough (Alaska) and the National Oceanic and Atmospheric Administration (through the Alaska Eskimo Whaling Commission) for supporting and funding this research.

Also, thanks to Dr. Michael F. Antolin, Department of Biology, for providing the black-tailed prairie-dog dataset and his expertise about population genetics.

Last but not least, I would like to thank my family for their endless support in achieving my goals in life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction to Genetics & Population Structure

1.1 Introduction to Genetics

Each cell in our body contains genetic information that is crucial to our survival and diversity. In my dissertation, I focus on organisms that are diploid, i.e., have chromosomes that are made up of two copies of double stranded molecules called DNA (deoxyribonucleic acid). The number of chromosomes varies from one species to the other. While humans have 23 pairs of chromosomes, whales have 22 and black-tailed prairie dogs have 25 pairs of chromosomes [43]. DNA can be in the nucleus of a cell or in the mitochondria (mtDNA). Mitochondrial DNA carries genetic information inherited only from the mother and does

not carry any information from the father. MtDNA is therefore haploid because it consists of only one copy of a gene. In my dissertation I focus on genetic information collected from the DNA in the nucleus, i.e., nuclear DNA.

DNA carries genetic coding for the existence and inheritance of the species. Genes are segments of the DNA as shown in Figure 1.1. The location in the chromosome where a gene is located is called a locus. Any specific gene may exhibit several different forms; each type is called an allele. Thus, each locus within an individual consists of a pair of alleles each located on one of the chromosomes. Figure 1.2 displays two chromosomes from a flower's cell nucleus. The gene of interest is the color of the flower. There are various possible color types the alleles can take. Figure 1.2 displays the case where one allele is purple flower and the other is white.

When one of the gene types is dominant over another, it is called a dominant gene.

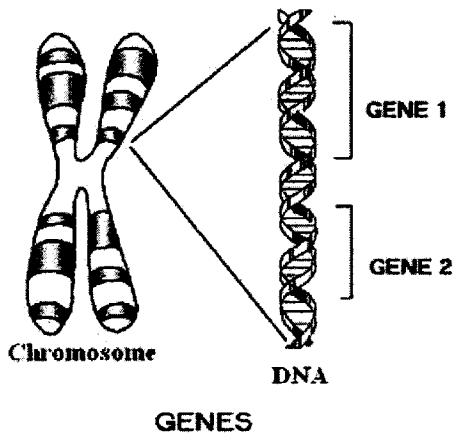

Figure 1.1: Figure of chromosomes and its components for a diploid organism [19].

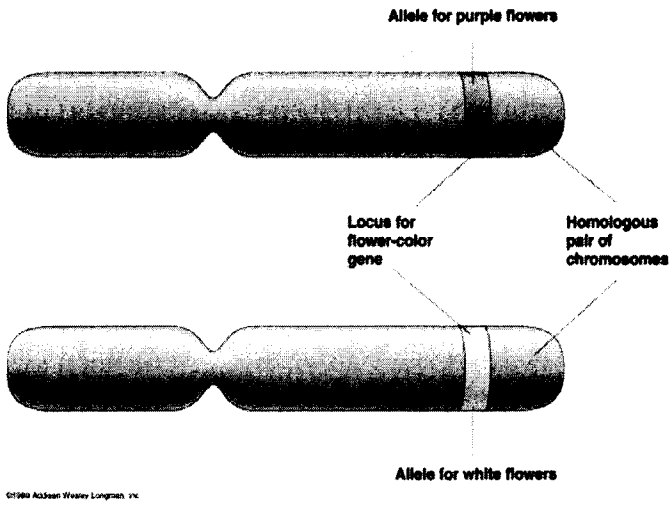

Figure 1.2: Figure of a chromosome of a flower. The locus is where the gene determining the color of the flower is located. White and purple are two of the various alleles for that specific locus [15].

Similarly the gene that is being dominated is called the recessive gene. In this case, the genetic information that the dominant gene carries is manifested in the organism phenotype while the recessive gene is hidden. As an example, someone with blood type A may have allele pair (or genotype) AA or AO. Since A is a dominant gene and O is a recessive gene, the person displays the property of gene type A, i.e., he has phenotype A. Therefore, it is not always possible to determine the types of genes that an individual has by looking at the phenotype.

When the two alleles at a certain locus are of the same type (AA in the blood type example), the genotype is said to be homozygous and when they are different (AO in the blood type example) the genotype is said to be heterozygous. One of these alleles is inherited from the mother while the other is inherited from the father.

The method that I develop in this dissertation can be applied to various different types of genetic data. Here, I will concentrate my applications on a genetic data type called microsatellites. One of the molecule types DNA consists of are nucleotide bases. There are four different nucleotide bases; adenine (A), thymine (T), guanine (G) and cytosine (C) [36]. Figure 1.3 demonstrate this substructuring of the chromosome to the nuclear bases. Microsatellites consist of multiple sequential replications of (usually) two to four of these bases within a locus. For so-called variable microsatellites, the number of times these replications happen may differ between individuals of the same species. These numbers of replications, i.e., microsatellite alleles are heritable. Thus, microsatellites are used in a variety of different research areas. A survey of the methods and materials used to collect microsatellite data and the uses of such data is given by [58]. In forensic science, microsatellite data are used for DNA testing where the genetic data of the criminal and suspect are compared. In biomedical research microsatellites are used to help study medical diseases. In biological/evolutionary research, microsatellite data are crucial in answering questions concerning degree of relatedness of individuals and population structure.

The genome of each individual is made out of thousands of genes that determine how the organism functions. Each individual also passes on some of its genetic information to its offspring. There are various factors influencing which gene information is passed on and how this affects the genotype of the new individual. Individuals that breed only with members of their group can be thought of as a genetically distinct population. Such a population comprises a gene pool with various factors affecting its allele frequencies. It is possible — indeed normal— for a species to include multiple, genetically distinct subpopulations that

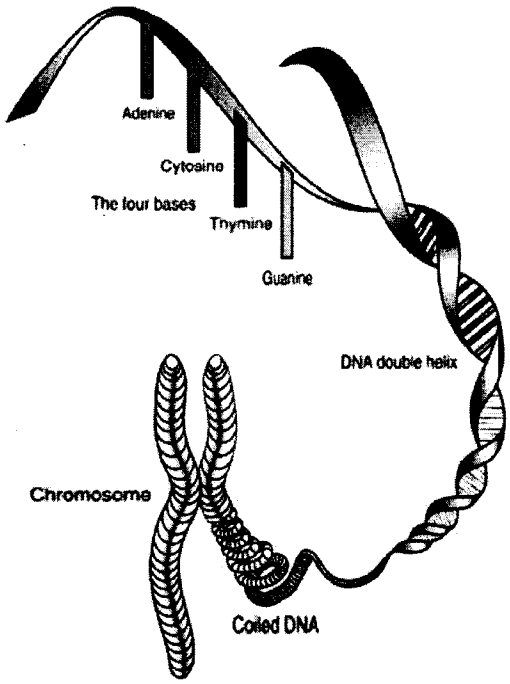

Figure 1.3: Decomposition of a chromosome to its nucleotide bases components for a diploid organism [3].

can be distinguished because these subgroups exhibit different population-specific allele frequencies (along with other genetic differences).

Inbreeding, subdivision of populations, and genetic flow are some of the factors influencing biological population structure. By analyzing genetic patterns and variation, population structure can be detected and explained. The existence of population structure means that individuals in the same subpopulation will be more genetically similar, on average, than individuals from distinct subpopulations. This tendency towards similarity is what I will informally refer to as genetic correlation. In this dissertation I will develop a new statistical method to detect genetic patterns of genetic correlation, particularly pat-

terns related to measurable covariates.

1.2 Population Structure and the Hardy-Weinberg Principle

The Hardy-Weinberg principle is one of the most fundamental concepts used in population genetics [45]. It states that after one generation of random mating, genotypic frequencies will follow a multinomial (binomial in biallelic case) distribution where the multinomial cell probabilities can be expressed as simple functions of the allelic frequencies. These cell probabilities can be described by the model that the alleles for any individual are chosen randomly and independently of each other. For Hardy-Weinberg equilibrium to hold, the following assumptions must hold for the population :

- Mutation is not occurring
- Natural selection is not occurring
- Population is infinitely large
- All members of the population breed
- All mating is totally at random
- Every individual in the population produces the same number of offspring
- There is no migration in or out of the population

Let X be the number of allele A in a genotype, i.e. $X = 0, 1$ and 2. For the biallelic case, with allele A frequency p_A, the Hardy-Weinberg principle states that the genotypic frequencies are:

$$\begin{aligned} P(X = 2|p_A) &= P(AA|p_A) = p_A^2 \\ P(X = 1|p_A) &= P(AB|p_A) = 2p_A(1 - p_A) \\ P(X = 0|p_A) &= P(BB|p_A) = (1 - p_A)^2 \end{aligned} \tag{1.1}$$

if the Hardy-Weinberg assumptions are true. If a population does not exhibit Hardy-Weinberg genotypic frequencies, then at least one of the assumptions is violated. This is important for detecting population structure because Hardy-Weinberg disequilibrium can be viewed as idealized single stock structuring. Due to various factors such as physical or ecological changes, with time, populations may get subdivided. For example, individuals may prefer to mate with those near by. Thus, the physical distance may have an impact in grouping within a population. If between-group mixing is sufficiently rare, the subgroups start differing in genetic patterns. Thus, if there is no mixing between subpopulations (or extremely little mixing), this can be detected by exploring the genotypic frequencies and their departure from Hardy-Weinberg proportions. For example, let a population be subdivided into two substocks having different allelic frequencies. Even if each subpopulation holds Hardy-Weinberg proportions, when the two subpopulations are lumped together, there will be a deficiency of heterozygotes and excess of homozygotes with respect to overall Hardy-Weinberg proportions. This is known as the Wahlund effect [56].

For illustration, suppose there exist two biallelic subpopulations of equal size, both under Hardy-Weinberg disequilibrium with allele type A frequencies, 0.4 and 0.8, respectively. The genotypic frequencies are calculated using (1.1) for each subpopulation and are given in the first two columns for Table 1.1.

Table 1.1: Genotypic frequencies for two subpopulations with allele type A frequencies 0.4 and 0.8, respectively. The final two columns are the genotype frequencies that would be expected and observed if the two populations were pooled.

	Subpopulation 1 ($p_{1A} = 0.4$)	Subpopulation 2 ($p_{2A} = 0.8$)	Expected ($\bar{p} = 0.6$)	Observed
AA	.16	.64	.36	.4
AB	.48	.32	.48	.4
BB	.36	.04	.16	.2

When the two subpopulations are lumped together, if Hardy-Weinberg proportions still hold, the expected genotypic frequencies are found using (1.1) where allele type A frequency is taken as the average of 0.4 and 0.8, namely 0.6. The expected frequencies for the pooled population are given in the third column of Table 1.1. The observed genotypic frequency when the two subpopulations are lumped together is the average of the genotypic frequencies of the subpopulations and is given in the last column of Table 1.1. Since each substock is under Hardy-Weinberg equilibrium, the difference between the expected and observed genotypic frequencies when the two substocks are lumped together is termed the Wahlund effect and signals substock structure.

There have been various methods developed to test for a Wahlund effect. These tests are based on whether or not the data reflect Hardy-Weinberg proportions. The simplest such test is a χ^2 test comparing the numbers of observed genotypes to expectations under Hardy-Weinberg.

Inbreeding can also cause departure from Hardy-Weinberg proportions. Also known as a form of non-random mating, inbreeding occurs when mating individuals are more closely related than those drawn at random from a subpopulation. The effect of inbreeding is a deficiency of heterozygotes and excess of homozygotes with respect to Hardy-Weinberg proportions. The coefficient of inbreeding, f, is defined as the probability that the two homologous alleles in an individual are identical by descent (ibd); i.e., alleles are both copies of one particular allele possessed by a common ancestor. It is also known as the correlation of alleles within individuals [7].

For biallelic locus, let the allele A frequency be p_A. Then the probability of two randomly chosen alleles both being allele type A can be expressed in terms of being ibd or not. That is, if the alleles are ibd then the corresponding probability is $p_A f$ while if they are not ibd the probability is $p_A^2(1-f)$. Putting these two together, the overall probability of an AA genotype is

$$\begin{aligned} P(AA) &= p_A^2 + f p_A(1-p_A) \\ &= p_A[f + (1-f)p_A]. \end{aligned} \tag{1.2}$$

This exceeds the proportion expected under Hardy-Weinberg, which is p_A^2. Thus, the non-zero inbreeding coefficient, f, results in an excess of homozygotes and a deficiency in heterozygotes.

Notice that the Wahlund effect and inbreeding both result in heterozygote deficiency. Therefore, one cannot use heterozygote deficiency alone to diagnose whether a population has substructure, inbreeding and/or some other patterns of genetic variation. The method I will describe in this dissertation enables one to separate various sources of excess homozygosity.

Let us take a closer look at genotype frequencies and genetic correlation in a more general case with k subpopulations. Let p_{iA} be the allele type A frequency at a biallelic locus in subpopulation i, for $i = 1, \cdots, k$. Define

$$\bar{p} = \frac{\sum p_{iA}}{k}. \tag{1.3}$$

Then, the expected AA genotypic frequency over all equally-sized subpopulations is

$$\begin{aligned} \frac{\sum p_{iA}^2}{k} &= \frac{\sum (p_{iA} - \bar{p} + \bar{p})^2}{k} \\ &= \bar{p}^2 + \frac{\sum (p_{iA} - \bar{p})^2}{k} \\ &= \bar{p}^2 + \mathrm{Var}(p) \\ &= \bar{p}^2 + \theta\bar{p}(1 - \bar{p}). \end{aligned} \tag{1.4}$$

Since Var(p) is always non-negative (equal to zero if all p_{iA} are the same), the result of lumping k subpopulations is more than the expected number of homozygotes and therefore, less than the expected number of heterozygotes. Comparing (1.2) and (1.4), the Wahlund effect, represented by θ in (1.4), influences genotypic frequency in the same direction as inbreeding, i.e., as an excess of homozygotes and deficiency of heterozygotes. But while inbreeding reduces heterozygote frequencies at all loci, the Wahlund effect influences only the heterozygote frequencies at the particular loci with allele frequency variation over subpopulations [45].

θ in equation (1.4) and f in equation (1.2) are also known as Wright's F statistics [61]. θ measures relatedness of pairs of alleles within a population relative to the total population and thus is often represented as F_{ST}. Similarly, f also known as the within population inbreeding coefficient or correlation structure within individuals relative to a population is represented by F_{IS} and the total inbreeding coefficient also defined as correlation structure within individuals relative to the total population is F_{IT}. The relationship between these three statistics is such that $(1 - F_{IT}) = (1 - F_{IS})(1 - F_{ST})$ [11, 12]. This equation can be derived using heterozygosity. Let H_I be the true heterozygosity of individuals within populations, H_S be the expected heterozygosity within subpopulations assuming Hardy-Weinberg within populations, and H_T be the expected heterozygosity in the combined population assuming Hardy-Weinberg over the whole sample. Then,

$$\begin{aligned} F_{IT} &= 1 - \frac{H_I}{H_T} \\ &= 1 - \frac{H_I}{H_S}\frac{H_S}{H_T} \\ (1 - F_{IT}) &= (1 - F_{IS})(1 - F_{ST}) \end{aligned}$$

Table 1.2: Genotype counts for the Isotoma petraea (bi-allelic) dataset provided by James et al. [53] for 8 populations. The last column represents allele A frequency in that particular population.

Population (i)	AA	AB	BB	allele A frequency (p_{iA})
1	14	3	3	0.7750
2	15	2	3	0.8000
3	13	0	0	1.0000
4	23	5	2	0.8500
5	23	3	4	0.8167
6	29	3	1	0.9242
7	5	0	0	1.0000
8	0	1	0	0.5000

[37].

In order to understand the application of Wright's F statistics, consider the Isotoma petraea (bi-allelic) dataset provided by James et al. [53]. The dataset is given in Table 5.1. Allele A frequency is found by taking the ratio of allele A counts and the total number of alleles in the population. For example, the allele A frequency in population 1 is found by $(14 * 2 + 3)/[2 * (14 + 3 + 3)] = 0.775$. If population allele frequencies were observed with no sampling error, then the three F statistics can be estimated with no bias using heterozygosity, sample estimate for $\bar{p}$ and sample variance of allele A frequency.

From Table 1.2, the mean of the allele A frequencies, using equation (1.3), gives $\bar{p} = 0.8332$ and $\text{Var}(p) = 0.02250$. Thus, using results in equation (1.4), $F_{ST} = \theta = 0.1619$. The observed individual heterozygosity is the average proportion of AB genotypes in the populations. Thus, $H_I = (0.150 + 0.100 + 0 + 0.167 + 0.100 + 0.091 + 0 + 1)/8 = 0.2009$. The expected heterozygosity is found using $\bar{p}$, i.e. $H_T = 2 * (0.8332)(1 - 0.8332) = 0.2779$. Thus, $F_{IT} = 1 - \frac{0.2009}{0.2779} = 0.2769$. Using equation (1.5), F_{IS} is solved and found to be 0.1372.

Notice in the example given above we assumed that there is no sampling error. In reality this is not the case resulting in bias due to such sampling error which can occur either when sampling only some individuals from a larger group of individuals within populations (statistical sampling) or when sampling some populations from a larger group of populations (genetic sampling) [7]. There has been various research done to develop estimates that take into account sampling error such as Nei's G_{ST} [38], Weir and Cockerham's θ [9, 10].

1.2.1 Testing Hardy-Weinberg Equilibrium

The genotypic frequency of AA is given in equation (1.1), when Hardy-Weinberg equilibrium holds and in equation (1.2) when it does not. The difference between these two probabilities is called the disequilibrium coefficient, D_A [7]. Thus, when Hardy-Weinberg equilibrium does not hold, equation (1.1) can be rearranged as

$$\begin{aligned} P(AA|p_A) &= p_A^2 + D_A \\ P(AB|p_A) &= 2p_A(1 - p_A) - 2D_A \end{aligned} \tag{1.5}$$

$$P(BB|p_A) = (1-p_A)^2 + D_A.$$

D_A can be thought of as a parameter measuring the deviation from Hardy-Weinberg equilibrium. Let P_{AA} be the genotypic frequency of AA in a biallelic setting. Then D_A can be estimated using the maximum likelihood estimator of $\hat{D}_A = \hat{P}_{AA} - \hat{p}_A$ where $\hat{P}_{AA}$ and $\hat{p}_A$ are sample estimates [7]. Similarly, the variance of the estimate $\hat{D}_A$ can be found.

Since the disequilibrium coefficient is a measure of departure from Hardy-Weinberg equilibrium, the null hypothesis for testing Hardy-Weinberg equilibrium is $H_O : D_A = 0$. There are various methods that have been used to test such hypotheses. I first start with methods with asymptotic distributional assumptions and continue with methods that do not make any asymptotic distributional assumptions.

Large Sample Approximation Methods

For large samples, assume that the MLE of D_A has a normal distribution with mean $E[\hat{D}_A]$ and variance $\text{Var}(\hat{D}_A)$. Then, the null hypothesis $H_O : D_A = 0$ is tested using the test statistic

$$Z = \frac{\hat{D}_A - E[\hat{D}_A]}{\sqrt{\text{Var}(\hat{D}_A)}} \tag{1.6}$$

Notice that testing for heterozygote deficiency is equivalent to testing $D_A < 0$ and similarly, testing for heterozygote excess is equivalent to testing $D_A > 0$. Therefore, Hardy-Weinberg equilibrium can be tested using a two-sided test where the alternative is Hardy-Weinberg

Table 1.3: Goodness-of-fit table for biallelic case. Expected genotypic frequencies are formulated under the null hypothesis that Hardy-Weinberg equilibrium holds.

	Observed	Expected	(Observed - Expected)^2/Expected
AA	n_{AA}	$n\hat{p}_A^2$	$(n\hat{D}_A)^2/n\hat{p}_A^2$
AB	n_{AB}	$2n\hat{p}_A\hat{p}_B$	$4(n\hat{D}_A)^2/2n\hat{p}_A\hat{p}_B$
BB	n_{BB}	$n\hat{p}_B^2$	$(n\hat{D}_A)^2/n\hat{p}_B^2$

equilibrium does not hold, or a one-sided test could be used to test either heterozygote deficiency or excess.

The square of a standard normal has a chi-squared distribution with 1 degree of freedom and can be used in testing Hardy-Weinberg equilibrium. Taking the square of (1.6) and using the large sample variance of $\hat{D_A}$, the chi-squared test statistic

$$Z^2 = \frac{n\hat{D}_A^2}{\hat{p}_A^2(1-\hat{p}_A^2)} \tag{1.7}$$

can be used for testing Hardy-Weinberg equilibrium. Due to the nature of the chi-square distribution, the alternative hypothesis is two-sided, i.e. $H_1 : D_A \neq 0$.

Another path to Z^2 in (1.7) is to construct a goodness-of-fit chi-squared test. Table 1.3 displays the observed number of genotypes, expected number of genotypes under Hardy-Weinberg equilibrium for a biallelic case where n_{AA}, n_{AB} and n_{BB} are the number of genotypes AA, AB and BB in the sample, respectively. We end up with the test statistic in (1.7) by summing the last column in Table 1.3 to get the goodness-of-fit chi-squared test

statistic.

Distribution-Free Methods

The testing methods thus far all rely on asymptotic distributional assumptions. From the last column in Table 1.3 and the test statistics in (1.6) and (1.7), it can be seen that a small expected allele frequency will have an increasing effect on the test statistic value. Especially when the sample size is not large, using continuous distribution for discrete count data causes problems. Unlike these tests with asymptotic distributional assumptions, there are methods to test Hardy-Weinberg equilibrium without any asymptotic distributional assumption. Exact tests for Hardy-Weinberg equilibrium and likelihood ratio tests are two such methods.

Exact Tests:

[24] derives the exact test based on probabilities of all possible heterozygotes given allele count and creates the rejection region of size α. These probabilities are based on the assumption that they can be expressed in terms of allele counts and the number of heterozygotes [24]. Let n_A be the number of allele A and x be the number of heterozygotes in the sample. Then [24] derives these conditional probabilities as

$$P(x|n_A) = \frac{n!n_A!(2n-n_A)!2^x}{[(n_A-x)/2]!x![n-(n_A+x)/2]!(2n)!} \tag{1.8}$$

where n is the sample size. Next, all possible samples are ordered with respect to their probabilities. The p-value of the sample is found by summing the probabilities of samples

that are less than some function of the observed sample. The null hypothesis $H_O : D_A = 0$ is rejected if the p-value is less than the significance level α.

As an example, I consider an example given in [7]. Table 1.4 gives all possible sample values, each with allele A count value of 19, for testing Hardy-Weinberg equilibrium using exact tests. The table displays all of the possible samples, where the first sample is the observed sample, ordered according to their corresponding conditional probability calculated using equation (1.8) with n_A and x as defined previously. Although the data are multiallelic, it has been simplified to a biallelic case for simplicity. The p-value for the

Table 1.4: Exact test for Hardy-Weinberg equilibrium for a sample of size 40 with allele A count of 19. The first row displays the observed sample.

AA	AB	BB	Probability	Cumulative Probability
9	1	30	0	0
8	3	29	0	0
7	5	28	0.0001	0.0001
6	7	27	0.0023	0.0024
5	9	26	0.0205	0.0229
0	19	21	0.0594	0.0823
4	11	25	0.0970	0.1793
1	17	22	0.2308	0.4101
3	13	24	0.2488	0.6589
2	15	23	0.3411	1.0000

exact test is found by summing the probability values of the possible samples with 1,3,5,7 and 9 heterozygotes. Thus, we reject the null hypothesis of Hardy-Weinberg equilibruim with a significance level of 0.0229.

Unlike chi-square tests, exact tests do not suffer when the expected allele frequency is small. Moreover, since the probabilities of observing genotypic counts do not depend on allele frequencies, Hardy-Weinberg inference using exact tests are not affected by population allele frequencies. For a single locus case, exact chi-square test can be found by using a contingency table as given in the example above. The multi-locus version requires permutation procedures to find the p-value for the test. The number of possible samples with same gene frequencies and sample size grows exponentially with the number of alleles. This makes the calculation of the p-value for the exact chi-square test almost impossible for real life cases.

[57] introduces two methods to use computer simulation to estimate the significance level for the exact test: a Monte Carlo method and Metropolis algorithm. The Monte Carlo method has the advantage that sample size can be predetermined so that the desired level of significance is achieved while limiting total computation time. The Markov Chain method has the advantage that Hardy-Weinberg probabilities of each table (i.e., for each locus) do not need to be calculated and therefore the time spent on each table does not depend on the size of the table. When the performances of these two methods are compared, the Monte Carlo method is better for small sample sizes with large number of alleles and the Markov Chain method is most useful for sparse data with large sample size [57].

[39] proposed an alternative method, namely the probability test based on exact non-parametric procedure. By using Markov Chain Monte Carlo to estimate the p-value for the exact test, this approach decreases the computation time required for datasets containing large number of individuals and/or loci with large number of alleles. Thus, this probability test is robust to the increase in number of alleles, subpopulations or sample size.

Likelihood Ratio Tests:

An alternative test for testing $D_A = 0$ is based on the log likelihood ratio statistic. The likelihood function is maximized under the unconstrained model (L_1) and under the model where the parameter of interest is constrained by the corresponding null hypothesis value (L_O). By taking the ratio of the two likelihoods, the likelihood ratio statistic λ is derived and can be approximated by chi-square distribution. That is, $-2\ln\lambda$ has chi-square distribution with 1 degree of freedom.

Let n_{ij} be the number of individual of genotype j in population sample i. Let $n_{i.}$ be the total sample size of the i^{th} sample. Define $n_{.j}$ as the total number of j genotypes across all samples and $n_{..} = \sum n_{i.}$. Then define the test statistic as

$$\begin{aligned} G^2 &= -2\ln\lambda \\ &= -2\ln(L_O/L_1) \\ &= -2\sum\sum n_{ij}\ln(\frac{n_{ij}n_{..}}{n_{i.}n_{.j}}). \end{aligned}$$

Then G^2 is the likelihood ratio test for multinomial proportions and is used for testing the null hypothesis of Hardy-Weinberg disequilibrium [50]. As in the case of exact tests, G^2 does not require population allele frequencies.

1.2.2 Detecting Population Structure

In the previous section I described different methods for testing Hardy-Weinberg equilibrium. Now I will focus on testing for substock structure. Similar to the previous section, exact tests from contingency tables can be used here also, but they cause similar problems when some cells have small expected counts under the null hypothesis of Hardy-Weinberg equilibrium. This problem becomes even more severe when the numbers of alleles increase. Therefore, in this section I will focus on methods that rely on the bootstrap, permutations, and so forth.

One of the most straightforward strategies for detecting population structure is to compare allele frequencies between strata. I begin with a discussion of several such techniques.

Bootstrap Methods:

Bootstrap methods can be used to determine if two strata have the same allele frequency. For a sample of size n, repeated samples of size n are randomly chosen with replacement. Allele frequency is estimated in each repeated sample. These estimates provide a distribution from which the confidence interval for the allele frequency in each strata can be constructed without the need of Hardy-Weinberg equilibrium assumption. By evaluating the confidence intervals for the allele frequency of the two strata, a decision about the population structure can be made. If there exists population structure, then the two confidence

intervals should not overlap [7]

Permutation Tests:

Another distribution-free method for comparing allele frequencies is based on permutation testing. Permutation tests are very useful especially when other methods fail due to the large number of possible pairs of samples or the existence of a rare allele resulting in small expected frequency in the contingency table.

Let n_1 and n_2 be the sample sizes of the two strata with corresponding allele A_i count n_{i1} and n_{i2}. If it is assumed that Hardy-Weinberg equilibrium hold, a global test based on all alleles is developed using joint probability of allele A_i counts in the two strata. Conditioning this joint probability to the total count of allele A_i, the conditional probability of allele A_i counts in each of the two strata under null hypothesis of equal allele frequencies is,

$$P(n_{i1}, n_{i2}|n_{i1}+n_{i2}) = \frac{(2n_1)!(2n_2)!}{2n_1+2n_2}\frac{\prod(n_{i1}+n_{i2})!}{\prod n_{i1}!\prod n_{i2}!} \tag{1.9}$$

After the total group of alleles is permuted, it is divided into samples with pre-determined sample sizes. The significance is determined by the proportion of conditional probabilities of the permuted allele counts, calculated using (1.9), that are less than that for the observed value [7].

AMOVA:

Elston *et al.* consider using analysis of variance concepts to break down genetic variation [49]. This approach is called AMOVA or analysis of molecular variance. In the case of no interaction among loci, the overall genetic variance is simply the sum of the single locus variances. Excoffier explains the hierarchical analysis of molecular variance from the matrix of squared distances between allele pairs [34]. This method is flexible since adjustment to assumptions can be done by adjusting the distance matrix. Thus, unlike ANOVA, normality assumption is not a requirement and thus AMOVA succeeds in explaining the hierarchy in the population structure.

For simplicity, I assume allele A is the allele of interest. Let x_{ij} be an indicator variable for allele A from population i. That is, $x_{ij} = 1$ if jth allele from population i is A and zero otherwise. A hierarchical analysis of variance breaks down the total variance into covariance components which are later used to estimate Wright's F statistics. Let the i^{th} allele frequency vector from j^{th} population be a linear equation of the form $x_{ij} = \mu + a_j$ where the vector μ is the unknown expectation of x_{ij} averaged over the whole study, and a is the population effect. Then the expectation of the binary variable x_{ij} is just the allele A frequency in population i. Thus testing for equal allele A frequencies in the populations can be done by applying ANOVA to the x_{ij}'s. Under the assumptions that the alleles in different populations are independent, Table 1.5 shows how the within-population and between-population correlation structure is broken down using x_{ij} for the k populations with n_i as the number of allele A in population i [7]. Note that $\hat{p}_{A.} = \frac{\sum n_i \hat{p}_{Ai}}{\sum n_i}$ and $\bar{p}_{A.} = \frac{\sum n_i p_{Ai}}{\sum n_i}$ where p_{Ai} is the true allele A frequency in population i and $\hat{p}_{Ai}$ is the sample

Table 1.5: ANOVA table for binary variable x_{ij} for $i = 1, \cdots, k$, $j = 1, \cdots, n_i$.

Source	df	Sum of Squares	Expected Mean Square
Among poulations	$k-1$	$\sum n_i(\hat{p}_{Ai} - \hat{p}_{A.})^2$	$\frac{1}{k-1}\sum(1 - \frac{n_i}{\sum n_i})p_{Ai}(1-p_{Ai})$ $+\frac{1}{k-1}\sum n_i(p_{Ai} - \bar{p}_{A.})^2$
Within populations	$\sum(n_i - 1)$	$\sum n_i\hat{p}_{Ai}(1-\hat{p}_{A.})$	$\frac{\sum(n_i-1)p_{Ai}(1-p_{A.})}{\sum(n_i-1)}$

allele A frequency in population i.

Clustering Methods:

Another strategy for population structure analysis is based on the statistical notion of clustering. Here, no group memberships or strata are specified. Instead the samples are analyzed to assess their nearness or genetic distance from each other. The dataset is partitioned empirically to form clusters that minimize within-cluster variation and maximize between-cluster variation in some sense. Ideally, Hardy-Weinberg disequilibrium is absent within each resultant cluster. The number of clusters may be fixed in advance or estimated. The process of forming clusters may be agglomerative, divisive (hierarchical) or model-based.

An example of such a method is the program Structure [13, 29]. Let the distribution of allele frequencies among populations be approximated by a Beta distribution with mean $\mathbf{p}$ and variance $\mathbf{p}(1-\mathbf{p})\theta$. Then by specifying priors on $\mathbf{p}$, θ and (unknown) population of origin for each individual (denoted by $\mathbf{Z}$), the posterior distribution can be used to estimate

the unknown parameters $\mathbf{p}$ and $\mathbf{Z}$. As a result of parameter estimation, Structure estimates ancestries or assignment probabilities of individuals to clusters. The number of clusters, K, must be fixed in advance. The likelihood that individual i comes from subgroup k can be found using Bayes' theorem. After specifying priors of genotypic frequencies, the posterior probability that individual i belongs to subgroup k can be estimated. Pritchard *et al.* [29] also suggest a likelihood-based way to compare choices for K, but the method is highly approximate and they warn against relying on this. [51] presents simulation results suggesting that the statistical power of Structure can be quite low when gene flow is moderate or high. The benefit of such clustering methods is that they focus more on assignment of individuals to subgroups, rather than on testing for the existence or number of subgroups.

1.3 New Method

The method described in this dissertation shares some similarities with many of the approaches discussed above. Like methods requiring pre-specification of groups, my method attempts to detect the signal of the stock structure rather than attempting to group individuals. Unlike many such methods, it associates genetic structure with covariates (along with an overall Wahlund effect). My method resembles Structure and similar techniques in that it is not reliant on a priori stratification of samples. The genetic signals from such groups should be detected without guessing or estimating group membership.

My method is based on empirical estimation of the similarity of paired alleles. All

alleles from all individuals are tested equally, and all pairwise comparisons are examined, both within and between individuals. I use covariates that may explain patterns in pairwise allele match probabilities. The significance of each term in my model is tested using permutation methods.

One of the datasets I use to illustrate the performance of my method relates to the Bering-Chukchi-Beaufort Seas population of bowhead whales. Population structure in this region is unknown, so genetic samples may come from one or more sub-populations (substocks). The stock identity of each whale is unknown. For this dataset, my goal is to accurately detect patterns of genetic variation among bowhead whales. These whales are migratory, and may exhibit spatio-temporal stock structure. Genetic samples are taken at only a few points along the migratory pattern path. One previous study, examining pairs of whales, has shown temporal genetic substructure [42]. Also, these whales are extremely long-lived and may have passed through a population bottleneck less than one whale lifetime ago [21]. Thus, the covariate "age" may be related to genetic structure caused by gene drift.

In Chapter 2, I develop a statistical model and estimation strategy useful in such applications. I also introduce a hypothesis testing strategy using permutation tests. In Chapter 3, I discuss simulation testing of my proposed model and compare the results to those from related methods. In the following chapter, Chapter 4, several applications of my method to real data are described. Some general conclusions and other discussions are given in Chapter 5.

Chapter 2

Modeling

This chapter describes the theoretical motivation for my modeling approach. I will build up the mathematics starting from the simplest case and moving to the most complex: First we will consider the single-locus bi-allelic case, then the single-locus multi-allelic case and finally the multi-locus multi-allelic case.

To investigate genetic stock structure, I will analyze the probability that two randomly selected alleles match. I will relate this probability to the identities and characteristics of the whale(s) from which the alleles were taken. A match occurs when two randomly sampled alleles are identical in state. All possible pairs of alleles in the sample are considered.

My model is based on allele match probabilities. Useful background on such probabilities (in the context of forensics) is given in Balding and Nichols [4] and in Ayres and Overall [33]. I will model these match probabilities using the same sorts of tools used for Generalized Linear Models, and Generalized Additive Models, using a covariate suspected

to be related to genetic variation among individuals and/or substocks.

2.1 Single Locus Bi-allelic Case

For simplicity, I will introduce the simplest case first: the single locus bi-allelic case. Here I assume there is only one locus with only two possible alleles. Although unrealistic compared to real life situations, this simplest case is intended to introduce the basic ideas of my methodological approach.

Assume there are at most two substocks. Assume there is one locus with alleles A and B. Let p_{iA} be the substock frequency of allele A from substock i, for $i = 1, 2$. Notice that since there are only two possible allele types, the substock frequency of allele B from substock i is $1 - p_{iA}$, for $i = 1, 2$. In the simplest case of two alleles chosen at random from a substock, with random mating within substocks,

$$P(AA|p_{iA}, \text{alleles from same substock}) = p_{iA}(\theta + (1 - \theta)p_{iA})) \tag{2.1}$$

where p_{iA} is the substock proportion of allele A and θ (defined as $1/F_{ST} - 1$ in [4]) is the probability of two alleles being identical by descent from a common ancestor in the same substock. This expression is easily generalized when there are more than two allele types at the locus. My models are intended for the case when θ may be greater than zero (because there is significant substock structure) and there is an association between a covariate of interest, X, and allele match probabilities.

When the random mating assumption is not valid,

$$P(AA|p_{iA}, \text{alleles from same substock}) = p_{iA}(f + (1-f)(\theta + (1-\theta)p_{iA})) \tag{2.2}$$

where f is a measurement of within substock correlation. Defining $g = f + \theta - f\theta$, it is easy to show that (2.2) reduces to

$$P(AA|p_{iA}, \text{alleles from same substock}) = p_{iA}(g + (1-g)p_{iA})). \tag{2.3}$$

Alleles drawn from separate source substocks are assumed to be independent. Thus, the probability of a match is

$$P(\text{match}|\text{alleles from different substocks}) = p_{1A}p_{2A} + (1-p_{1A})(1-p_{2A}). \tag{2.4}$$

In the initial stages of model development, I have taken the simplest case: assuming that the genetic data for each whale consist of a single locus with two allele types A and B. Also, I have assumed that whales originate from no more than two substocks, and that there is no migration between the two substocks. Of course, for the bowhead whales mentioned above, these assumptions may not hold. I now begin generalization of this allele matching probability model for a more complex situations that are more biologically reasonable.

2.1.1 Marginal Match Probabilities for Independent Substocks

Recall that the probability of observing allele A in substock i is p_{iA}, for $i = 1, 2$. Alleles are not independent when they are drawn from the same substock and allele match probabilities must account for the correlation. To develop my statistical model for match probabilities, I begin with the genetic probability model of Ayres [33] which views population

allele frequency parameters (the p_{iA}) as random variables which vary between substocks.

In developing my model, the probability of a match is marginalized over the distribution of substock allele frequencies. For example, consider the event AA. Let I_{AA} and I_A be the indicator that the allele pair is AA and that an allele is A, respectively. Then $I_{AA} = I_A I_A$. The marginal probability of an AA match is

$$\begin{aligned} P(AA\text{match}|\text{both alleles from substock } i) &= E(I_{AA}) = E(I_A I_A) \\ &= E_p(E[I_A I_A]|p_i) \\ &= E_p(gp_i + (1-g)p_i^2) \end{aligned} \tag{2.5}$$

where f, θ and g are as defined previously (i.e., $g = f + \theta - f\theta$). Assume the $p_i|p$ have independent beta distributions with mean p and variance $gp(1-p)$. Thus, marginally,

$$\begin{aligned} &P(AA\text{match}|\text{both alleles from substock } i) \\ &= gp + (1-g)(gp(1-p) + p^2) \\ &= p^2 + gp(1-p) + g(1-g)p(1-p). \end{aligned} \tag{2.6}$$

With a similar approach it can be shown that marginal probability of a BB match for alleles sampled from the same substock is

$$\begin{aligned} &P(\text{BB match}|\text{both alleles from substock } i) \\ &= (1-p)^2 + gp(1-p) + g(1-g)p(1-p). \end{aligned} \tag{2.7}$$

Therefore, the marginal probability of a match of any sort is the sum of (2.6) and (2.7). Thus, the marginal probability of a match when alleles are drawn from the same substock is:

$$P(\text{match}|\text{both alleles from same substock}) =$$

$$p^2 + (1-p)^2 + 2gp(1-p) + 2g(1-g)p(1-p). \tag{2.8}$$

Notice that if allele pairs are drawn from the same whale then they are also drawn from the same substock. Thus, the probability of a match when alleles are drawn from the same whale is equal to the probability of a match when alleles are drawn from same substocks. Thus, (2.8) can be thought of as the match probability for the case when allele pairs are drawn from the same whale.

I need to develop a model that does not depend on the knowledge of which substock is the source for each allele, since this information is not available in most microsatellite datasets and the existence of a separate second substock is uncertain. To do this note that the probability that two alleles from different whales match is,

$$\begin{aligned}
&P(\text{match}|\text{alleles from different whales}) \\
&= P(\text{alleles from same substock}) \\
&* P(\text{match}|\text{alleles from different whales from same substock}) \\
&+ P(\text{alleles from different substock}) \\
&* P(\text{match}|\text{alleles from different whales in different substock}) \\
&= U(p^2 + (1-p)^2 + 2gp(1-p) + 2g(1-g)p(1-p)) + (1-U)(p^2 + (1-p)^2) \\
&= p^2 + (1-p)^2 + 2U[gp(1-p) + g(1-g)p(1-p)], \qquad (2.9)
\end{aligned}$$

where U is the probability that two sampled alleles are from the same substock. Equation (2.9) can be viewed as the weighted average of the probability of a match when both alleles are from the same substock and probability of a match when they are from different substocks with weights U and $(1-U)$, respectively.

2.1.2 Model for Detecting Stock Structure from Match Probabilities

Let Y be a binary variable indicating that two sampled alleles match, and let Y_{ij} be the value of Y for the allele pair consisting of alleles i and j. Then $Y_{ij} = 1$ if the ith and jth alleles match and zero otherwise. The distribution of the random variable Y_{ij} is Bernoulli with a parameter that depends on whether the alleles are drawn from the same whale or from two different whales. From (2.8), if the alleles are from the same whale, then Y_{ij} can be modeled as Bernoulli$(p^2 + (1-p)^2 + 2gp(1-p) + 2g(1-g)p(1-p))$. From (2.9), if the alleles are from two different whales, then Y_{ij} can be modeled as Bernoulli$(p^2 + (1-p)^2 + 2Ugp(1-p) + 2Ug(1-g)p(1-p))$.

Let $Q_{ij} = P(Y_{ij} = 1)$ and

$$Z_{ij} = \frac{Q_{ij} - p^2 - (1-p)^2}{2p(1-p)}. \tag{2.10}$$

Then,

$$Z_{ij} = \begin{cases} g + g(1-g), & \text{if } i\text{th and } j\text{th alleles are from same whale} \\ Ug + Ug(1-g), & \text{if alleles are from different whales.} \end{cases} \tag{2.11}$$

Suppose that genetic correlation varies smoothly with a covariate variable, X, within each substock, but the covariate does not induce correlation between alleles from separate substocks. Let X_{ij} be the value of the covariate measured for the ijth allele pair. For the bowhead data, the most important covariate is Δt_{ij}, the temporal separation of the capture times for ith and jth sampled alleles. Other potential covariates are the age difference and the length difference of the whale(s) providing the two alleles. For simplicity, I currently consider $X_{ij} = \Delta t_{ij}$ with $\Delta t_{ij} = 0$ if the ith and jth sampled alleles originate from the

same whale.

In order to motivate the class of models I propose, I begin with the assumption that f is linear in the covariate, so the value for a specific allele pair can be written as $f(X_{ij}) = \alpha_0 + \alpha_1 X_{ij}$. This assumption is biologically implausible and not necessary for my approach, but it is the simplest way to explain the idea behind my model. Later, I will generalize this so $f(X_{ij}) = s(X_{ij})$ for some smooth function s. The effect of assuming that f depends on X_{ij} is that the value of g in, e.g., equation (2.11) differs for each i and j. Specifically, defining $g(X_{ij}) = \theta + f(X_{ij}) - \theta f(X_{ij})$, I can re-express Z as

$$Z_{ij} = 2g(0) + 2(Ug(X_{ij}) - g(0))\delta_{ij} - g(0)^2 + (g(0)^2 - Ug(X_{ij})^2)\delta_{ij}, \tag{2.12}$$

where $\delta_{ij} = 1$ if alleles i and j are from different whales and zero otherwise.

Substituting $\mu = [g(0) + g(0)(1 - g(0))] = [2\alpha_0 - \alpha_0^2]$, $\gamma_1 = (U-1)\mu$, $\gamma_2 = 2U\alpha_1(1-\theta)(1-\alpha_0)$ and $\gamma_3 = -U\alpha_1^2(1-\theta)^2$, equation (2.12) can be simplified to

$$Z_{ij} = \mu + \gamma_1\delta_{ij} + \gamma_2 X_{ij}\delta_{ij} + \gamma_3 X_{ij}^2\delta_{ij}. \tag{2.13}$$

Evaluating (2.13) at $X_{ij} = 0$ gives the Z_{ij} value when allele pairs come from different whales and have covariate value zero, i.e., $\mu + \gamma_1$. If allele pairs come from the same whale, then $\delta_{ij} = 0$ and so $Z_{ij} = \mu$. If there is only one substock, then the model should give the same results whether allele pairs are from the same whale or different whales. In this case, $\mu + \gamma_1 = \mu$. Thus, testing $\gamma_1 = 0$ becomes equivalent to testing the hypothesis that the data originate from a single population.

Clearly γ_2 and γ_3 jointly quantify the effect of X on the allele match probability. The

effect of the covariate on f is investigated by testing the null hypothesis of $\gamma_2 = \gamma_3 = 0$.

The discussion above is reminiscent of a generalized linear model. However, the outcomes are not independent so inference cannot be based on the standard large-sample methods for Generalized Linear Models. I use permutation tests in implementing both of the hypotheses of interest. Details of hypothesis testing are given in later sections of this chapter.

2.1.3 Dependent Substocks Case

In developing the model in (2.13), it was assumed that the covariate of interest, X_{ij}, influenced the match probability only when alleles originated from the same substock. For many covariates, this is probably sensible. As an alternative, let us consider the case when the covariate also affects the allele pair match probability when alleles originate from different substocks.

Then, alleles drawn from different substocks are no longer independent. For alleles originating from different substocks with allele A frequencies p_{1A} and p_{2A}, respectively, the AA match probability is,

$$
\begin{aligned}
& P(AA\text{match}|\text{alleles from different substocks}) \\
= \;& P(s_i = 1)P(1_A(i){=}1|s_i = 1)P(1_A(j){=}1|s_j = 2) \\
+ \;& P(s_i = 2))P(1_A(i){=}1|s_i = 2)P(1_A(j){=}1|s_j = 1) \\
= \;& \frac{n_1}{n_1 + n_2}(p_{1A}(f + (1-f)p_{2A})) + \frac{n_2}{n_1 + n_2}(p_{2A}(f + (1-f)p_{1A})) \qquad (2.14)
\end{aligned}
$$

where n_1 and n_2 are the sample sizes of substock 1 and substock 2, respectively. $1_A(i)$ is the indicator that the ith drawn allele is allele type A, and s_i indexes the substock from which the allele is drawn. Therefore, the marginal probability of an AA match when alleles originate from different substocks is

$$E[P(AA\text{match}|\text{alleles from different substocks})] = p(f + (1-f)p) \tag{2.15}$$

since the p_i are independent with mean p.

Similarly, the marginal probability of a BB match when alleles originate from different substocks is,

$$E[P(\text{BB match}|\text{alleles from different substocks}] = (1-p)(f + (1-f)(1-p)) \tag{2.16}$$

Thus, the marginal probability of a match when alleles originate from different substocks is the sum of (2.15) and (2.16):

$$E[P(\text{match}|\text{alleles from different substocks}] = p^2 + (1-p)^2 + 2p(1-p)f. \tag{2.17}$$

Notice that the marginal probability of a match when alleles originate from the same substock is not affected by the new assumption. Thus equation (2.8) still holds and is valid for match probabilities when the paired alleles originate from the same whale.

Of course the source substock for each allele is still unknown, but the source whales are known. We seek an expression for the marginal match probability conditional on the source whales. To find this, let U be the probability of two randomly drawn alleles being from the same substock. Then the marginal probability of a match when two alleles are

randomly sampled from different whales is,

$$\begin{aligned} & P(\text{match}|\text{alleles from different whales}) \\ &= U(p^2+(1-p)^2+2gp(1-p)+2g(1-g)p(1-p)) \\ &+ (1-U)(p^2+(1-p)^2+2p(1-p)f) \\ &= p^2+(1-p)^2+2p(1-p)[U(g+g(1-g))+(1-U)f]. \end{aligned} \tag{2.18}$$

Let $Y_{ij} = 1$ if alleles i and j match, and $Y_{ij} = 0$ otherwise. From (2.18), Y_{ij} is distributed Bernoulli$(p^2+(1-p)^2+2gp(1-p)+2g(1-g)p(1-p))$ if the alleles are drawn from the same whale. If the alleles are drawn from different whales, Y_{ij} is distributed Bernoulli$(p^2+(1-p)^2+2p(1-p)[U(g+g(1-g))+(1-U)f])$. Let $Q_{ij} = P(Y_{ij}=1)$ and define

$$Z_{ij} = \frac{Q_{ij}-p^2-(1-p)^2}{2p(1-p)}. \tag{2.19}$$

Then

$$Z_{ij} = \begin{cases} g+g(1-g), & \text{if } i\text{th and } j\text{th alleles are from same whale} \\ Ug+Ug(1-g)+(1-U)f, & \text{if alleles are from different whales.} \end{cases} \tag{2.20}$$

Let us continue to assume a linear relationship between f and X, where $f = \alpha_0+\alpha_1 X_{ij}$ and X is the covariate of interest over a range of (0,1). As before, $g = \theta+f-f\theta$. Then I can arrange Z as,

$$Z_{ij} = 2g(0)+2(Ug(X_{ij})-g(0))\delta_{ij}-g(0)^2+(g(0)^2-Ug(X_{ij})^2)\delta_{ij}+(1-U)f\delta_{ij} \tag{2.21}$$

where $\delta_{ij} = 1$ if alleles i and j are from different whales and $\delta_{ij} = 0$ otherwise. After reparameterization, (2.21) simplifies to

$$Z_{ij} = \mu+\gamma_1\delta_{ij}+\gamma_2 X_{ij}\delta_{ij}+\gamma_3 X_{ij}^2\delta_{ij} \tag{2.22}$$

where $\mu = [2g(0) - g(0)^2]$, $\gamma_1 = (1-U)[\alpha_0 - 2\alpha_0^2 + f]$, $\gamma_2 = 2U\alpha_1(1-\theta)(1-\alpha_0)$, and $\gamma_3 = -U\alpha_1^2(1-\theta)^2$. With the same reasoning as in the independent substocks case, testing $H_o : \gamma_1 = 0$ corresponds to testing for the existence of more than one population, and testing $H_o : \gamma_2 = \gamma_3 = 0$ amounts to testing whether the covariate X affects f.

The important aspect of (2.22) is that from a different set of assumptions, the same model is derived. Therefore it is unnecessary to determine or to make assumptions about whether the covariate operates across substocks. The same modeling approach can be used in either case.

To summarize, I have developed two approaches in developing a model for the binary random variables Y_{ij} which equal 1 when alleles i and j match and zero otherwise. The model asserts that $Y_{ij} \sim$Bernoulli(Q_{ij}) so $E(Y_{ij}) = Q_{ij}$. I have a link function,

$$g(Q) = \frac{Q - p^2 - (1-p)^2}{2p(1-p)} \tag{2.23}$$

and the linear predictor $g(Q_{ij}) = \mu + \gamma_1\delta_{ij} + \gamma_2\delta_{ij}X_{ij} + \gamma_3(X_{ij})^2\delta_{ij}$. Together these are the components of a generalized linear model in the binomial family with logit link function [44]. Such a model can be fit easily in statistical packages such as S-Plus [35], [59]. Appropriate tests of $\gamma_1 = 0$ and $\gamma_2 = \gamma_3 = 0$ can be used, respectively, to test the multiple-substock hypothesis and the hypothesis that genetic similarity depends on the covariate X.

2.1.4 Hypothesis Testing in Single Locus Bi-allelic Case

Before proceeding to derive the model under more complex genetic assumptions, let us first explore hypothesis testing a little. I will construct permutation tests for testing the single substock hypothesis and for testing the covariate effect. Recall from previous sections that the model has a linear predictor given by

$$Z_{ij} = \mu + \gamma_1\delta_{ij} + \gamma_2 X_{ij}\delta_{ij} + \gamma_3 X_{ij}^2\delta_{ij}.$$

When $X_{ij} = 0$, allele pairs coming from same whales and from different whales have Z_{ij} values of μ and $\mu + \gamma_1$, respectively. If there actually is only one substock then γ_1 should be zero. If there exists two substocks then the allele match probability for different whales depends on whether the two paired alleles originate from the same substock or from different substocks. Marginally, therefore, the mixture of different gene pools decreases the match probability when compared with the case of single substock match probabilities. This can also be algrebratically explained: Recall that Y_{ij} ~Bernoulli (Q_{ij}) and U is the probability of two randomly drawn alleles being from the same substock, $0 \leq U \leq 1$. When Q_{ij} values are compared for alleles originating from the same whale given in (2.8) and alleles originating from different whales given in (2.9), it can be seen that the match probability (or Q_{ij}) for alleles originating from same whale is larger than those originating from different whales. Thus, it is expected that $\mu > \mu + \gamma_1$. Therefore, $\gamma_1 < 0$ could be considered as the alternative hypothesis.

Under the null hypothesis of single substock, shuffling the δ_{ij} with respect to the Y_{ij} is permissible because the δ_{ij} have no effect on the match probabilities. Thus, permutation

tests can be constructed by using the null distribution of γ_1 estimates or the null distribution of the deviance penalty attributable to omitting γ_1 from the model, where these null distributions are obtained by repeated shuffles of the δ_{ij}. If the original estimate for γ_1 or its associated deviance is significantly different from those obtained from the reshuffled datasets, then the null hypothesis is rejected.

Testing the covariate effect is equivalent to testing $\gamma_2 = \gamma_3 = 0$. That is, if X is not statistically significant then the match probabilities will not be affected by the portion of the model that depends on the X_{ij}. Thus, testing $\gamma_2 = \gamma_3 = 0$ is analogous to the temporal pattern test in Jorde et al. [42]. Similar to permutation tests for γ_1, permutation tests for covariate effect can be constructed by shuffling the X_{ij} with respect to the Y_{ij}. Under the null hypothesis of no X effect, this shuffling is permissible since the X_{ij} are unrelated to the match probabilities. Such shuffling can be used to generate a null distribution for the deviance penalty associated with omitting terms involving X from the model. If the observed deviance change is significantly different from that of the reshuffled datasets, then the null hypothesis is rejected and significant covariate effect is concluded.

2.2 Multi-Allele Single-Locus Case

Thus far, we have considered only the bi-allelic case. In reality, there are usually many more possible alleles for each locus. Denote the probability of observing allele type A_j in substock i as p_{ij}, for $i = 1, 2$ and $j = 1, \cdots, m$, where m denotes the total number of alleles

at the locus. The probability that two alleles chosen at random from the ith substock match is

$$P(A_jA_j|p_{ij}, \text{both alleles from } i\text{th substock}) = p_{ij}(f + (1-f)(\theta + (1-\theta)p_{ij})) \quad (2.24)$$

where f is a measurement of within substock correlation, p_{ij} is the ith frequency of allele A_j in the ith substock, and θ is as defined before. Assuming $g = f + \theta - f\theta$, (2.24) reduces to

$$P(A_jA_j|p_{ij}, \text{both alleles from } i\text{th substock}) = p_{ij}(g + (1-g)p_{ij})). \quad (2.25)$$

Alleles drawn from separate source substocks are assumed to be independent and therefore

$$P(\text{match}|\text{alleles from different substocks}) = \sum_{j=1}^{m} p_{1j}p_{2j}. \quad (2.26)$$

2.2.1 Marginal Match Probabilities for Independent Substocks

First we address the case when the covariate is assumed not to operate across substocks. Let $I_{A_jA_j}$ and I_{A_j} be the indicator that the allele pair is A_jA_j and that the allele is A_j, respectively. Then $I_{A_jA_j} = I_{A_j}I_{A_j}$. The marginal probability of an A_jA_j match is:

$$\begin{aligned} P(A_jA_j\text{match}|\text{both alleles from substock } i) &= E(I_{A_jA_j}) \\ &= E(I_{A_j}I_{A_j}) \end{aligned}$$

$$= E_p(E[I_{A_j} I_{A_j}]|p_i)$$
$$= E_p(gp_{ij} + (1-g)p_{ij}^2)$$

where $g = f + \theta - f\theta$, f and θ are as defined before. The $p_{ij}|p_j$ are independent with mean p_j and variance $gp_j(1-p_j)$. Thus, marginally,

$$\begin{aligned} & P(A_jA_j\text{match}|\text{both alleles from substock } i) \\ &= gp_j + (1-g)(gp_j(1-p_j) + p_j^2) \\ &= p_j^2 + gp_j(1-p_j) + g(1-g)p_j(1-p_j) \end{aligned} \tag{2.27}$$

A match is defined as two randomly drawn alleles being the same. Therefore, the marginal probability of a match is the sum of (2.27) over all possible values of j, for $j = 1, \cdots, m$. Thus, the marginal probability of a match when alleles are drawn from the same substock is

$$\begin{aligned} P(\text{match}|\text{both alleles from same substock}) &= \sum_{j=1}^{m} [p_j g + (1-g)[gp_j(1-p_j) + p_j^2]] \\ &= g + (1-g)g + (1-g)^2 \sum_{j=1}^{m} p_j^2. \end{aligned} \tag{2.28}$$

The marginal probability of a match when two alleles are drawn from different whales is

$$\begin{aligned} & P(\text{match}|\text{alleles from different substock}) \\ &= \sum_{j=1}^{m} P(A_jA_j\text{match}|\text{alleles from different substock}) \\ &= \sum_{j=1}^{m} E[p_{1j}p_{2j}] \\ &= \sum_{j=1}^{m} p_j^2 \end{aligned} \tag{2.29}$$

since the p_{ij} are independent with mean p_j and variance $gp_j(1-p_j)$.

In order to develop a model that does not depend on the knowledge of which substock is the source for each allele, define U as the probability that two sampled alleles are from the same substock. Then,

$$
\begin{aligned}
& P(\text{match}|\text{alleles from different whales}) \\
= \; & P(\text{alleles from same substock}) \\
* \; & P(\text{match}|\text{alleles from different whales from same substock}) \\
+ \; & P(\text{alleles from different substock}) \\
* \; & P(\text{match}|\text{alleles from different whales in different substock}) \\
= \; & U[g + g(1-g) + (1-g)^2 \sum_{j=1}^{m} p_j^2] + (1-U) \sum_{j=1}^{m} p_j^2 \\
= \; & \sum_{j=1}^{m} p_j^2 + U[g + g(1-g) - 2g \sum_{j=1}^{m} p_j^2 + g^2 \sum_{j=1}^{m} p_j^2] \\
= \; & \sum_{j=1}^{m} p_j^2 + U[2g(1 - \sum_{j=1}^{m} p_j^2) - g^2(1 - \sum_{j=1}^{m} p_j^2)]. \qquad (2.30)
\end{aligned}
$$

Equation (2.30) can be viewed as the weighted average of the probability of a match when both alleles are from the same substock and probability of a match when they are from different substocks with weights U and $(1-U)$, respectively.

2.2.2 Model for Detecting Stock Structure from Match Probabilities

Let Y be the binary variable indicating that two sampled alleles match, and Y_{ij} be the value of Y for the allele pair consisting of alleles i and j. Then, $Y_{ij} = 1$ if the ith and jth alleles match and zero otherwise. The distribution of the random variable Y_{ij} is Bernoulli with a parameter that depends on f, θ, and whether the alleles are drawn from the same

whale or from two different whales. Using the same interpretation as before, the probability of a match when alleles are drawn from the same whale is equivalent to the probability of a match when alleles are drawn from same substocks. From (2.28), if the alleles are from the same whale then Y_{ij} can be modeled as Bernoulli$(g+(1-g)[g+(1-g)\sum p_j^2])$. From (2.30), if the alleles are from two different whales, then Y_{ij} can be modeled as Bernoulli$(\sum_{j=1}^{m} p_j^2 + U[2g(1-\sum_{j=1}^{m} p_j^2) - g^2(1-\sum_{j=1}^{m} p_j^2)])$. Let $Q_{ij} = P(Y_{ij}=1)$ and

$$Z_{ij} = \frac{Q_{ij} - \sum_{j=1}^{m} p_j^2}{1 - \sum_{j=1}^{m} p_j^2} \tag{2.31}$$

Then, $Z_{ij} = g+g(1-g)$ if ith and jth alleles are from same whale and $Z_{ij} = Ug+Ug(1-g)$ otherwise.

When this link function was used for fitting the model, I encountered numerical difficulties. Estimates of p_j fell outside of the interval of $[0,1]$. Thus, for numerical practicality, I propose using the logit link function:

$$Z_{ij} = \log(\frac{Q_{ij}}{1-Q_{ij}}). \tag{2.32}$$

Notice that there is no need to estimate p_j, for $j = 1, \cdots, m$ where m is the number of alleles.

Suppose that genetic correlation varies smoothly with a covariate variable, X within each substock, but the covariate does not induce correlation between alleles from separate substocks. Let X_{ij} be the value of the covariate measured for the ijth allele pair. For the bowhead data, one important covariate is Δt_{ij}, the temporal separation of the ith and jth sampled alleles. Other potential covariates are the age difference and the length difference

of the whale(s) providing the two alleles. For simplicity, I currently consider $X_{ij} = \Delta t_{ij}$.

In order to motivate the class of models I propose, I again begin with the assumption that f is linear in the covariate, so the value for a specific allele pair can be written as $f(X_{ij}) = \alpha_0 + \alpha_1 X_{ij}$. This assumption is biologically implausible and not necessary for my approach, but it is the simplest way to explain the idea behind my model. Later, I will generalize this so $f(X_{ij}) = s(X_{ij})$ for some smooth function s. Defining $g(X_{ij}) = \theta + f(X_{ij}) - \theta f(X_{ij})$, I can re-express Z as,

$$Z_{ij} = 2g(0) + 2(Ug(X_{ij}) - g(0))\delta_{ij} - g(0)^2 + (g(0)^2 - Ug(X_{ij})^2)\delta_{ij}, \tag{2.33}$$

where $\delta_{ij} = 1$ if alleles i and j are from different whales and zero otherwise.

Substituting $\mu = [g(0) + g(0)(1 - g(0))] = [2\alpha_0 - \alpha_0^2]$, $\gamma_1 = (U-1)\mu$, $\gamma_2 = 2U\alpha_1(1 - \theta)(1 - \alpha_0)$ and $\gamma_3 = -U\alpha_1^2(1-\theta)^2$, equation (2.33) can be simplified to

$$Z_{ij} = \mu + \gamma_1\delta_{ij} + \gamma_2 X_{ij}\delta_{ij} + \gamma_3 X_{ij}^2\delta_{ij}. \tag{2.34}$$

When (2.34) is evaluated at $X_{ij} = 0$, allele pairs coming from same whale and pairs coming from different whales have Z_{ij} values of μ and $\mu + \gamma_1$, respectively. The concept of hypothesis testing in previous sections remains valid here. If there actually is only one substock then γ_1 should be zero. If there exists two substocks then the allele match probability for different whales consists of match probability for different whales from same substock and that for from different substocks. The mixture of different gene pools decreases the match probability when compared with the case of single substock match probabilities. Thus, testing one substock versus the alternative hypothesis of more than one substock (in our case 2 substocks) is equivalent to testing $\gamma_1 = 0$.

Clearly γ_2 and γ_3 jointly quantify for the effect of X on allele match probability. The effect of the covariate on f is tested by testing the null hypothesis of $\gamma_2 = \gamma_3 = 0$.

2.2.3 Marginal Match Probabilities for Dependent Substocks Case

Now consider the case when the covariate also affects the allele pair match probability when alleles originate from different substocks. In this case, alleles drawn from different substocks are no longer independent. For alleles originating from different substocks with allele A_j frequencies p_{1j} and p_{2j}, the A_jA_j match probability is

$$\begin{aligned}
&P(A_jA_j\text{match}|\text{alleles from different substocks}) \\
&= P(\text{pop=1})P(1_{A_j}\text{=1}|\text{pop=1})P(1_{A_j}\text{=1}|\text{pop=2}) \\
&+ P(\text{pop=2})P(1_{A_j}\text{=1}|\text{pop=2})P(1_{A_j}\text{=1}|\text{pop=1}) \\
&= \frac{n_1}{n_1+n_2}(p_{1j}(f+(1-f)p_{2j})) \\
&+ \frac{n_2}{n_1+n_2}(p_{2j}(f+(1-f)p_{1j})) \\
&= \frac{n_1}{n_1+n_2}p_{1j}f + \frac{n_2}{n_1+n_2}p_{2j}f + (1-f)p_{1j}p_{2j}, \qquad (2.35)
\end{aligned}$$

where n_1 and n_2 are the sample sizes from substock 1 and substock 2, respectively, 1_{A_j} is the indicator that the allele is A_j, and pop defines the substock the allele is drawn. Therefore, the marginal probability of an A_jA_j match when alleles originate from different substocks is

$$E[P(A_jA_j\text{match}|\text{alleles from different substocks})]$$

$$
\begin{aligned}
&= E[\frac{n_1}{n_1+n_2}p_{1j}f + \frac{n_2}{n_1+n_2}p_{2j}f + (1-f)p_{1j}p_{2j}] \\
&= p_j(f+(1-f)p_j), \qquad (2.36)
\end{aligned}
$$

since the p_{ij} are independent with mean p_j.

Thus the marginal probability of a match when alleles originate from different substocks is the sum of (2.36) over all possible values of j, for $j = 1, \cdots, m$. Thus,

$$
\begin{aligned}
E[P(\text{match}|\text{alleles from different substocks})] &= \sum_{j=1}^{m}[p_j(f+(1-f)p_j)] \\
&= f + (1-f)\sum_{j=1}^{m} p_j^2. \qquad (2.37)
\end{aligned}
$$

Notice that the marginal probability of a match when alleles originate from the same substock (and thus same whale) is not affected by the new assumption and so equation (2.28) still holds.

Let U be the probability of two randomly drawn alleles being from same substock. From (2.28) and (2.37), we find that the marginal probability of a match when two alleles are randomly sampled from different whales is

$$
\begin{aligned}
P(\text{match}|\text{alleles from different whales}) &= U(g+(1-g)[g+(1-g)\sum_{j=1}^{m} p_j^2]) \\
&+ (1-U)[\sum_{j=1}^{m} p_j f + (1-f)p_j^2] \\
= \sum_{j=1}^{m} p_j^2 + U[g+g(1-g)](1-\sum_{j=1}^{m} p_j^2) &+ (1-U)f(1-\sum_{j=1}^{m} p_j^2). \qquad (2.38)
\end{aligned}
$$

Equation (2.38) can be viewed as the weighted average of the probability of a match when both alleles are from the same substock and probability of a match when they are from different substocks with weights U and $(1-U)$, respectively.

2.2.4 Model for Detecting Stock Structure from Match Probabilities

Let $Y_{ij} = 1$ if alleles i and j match, and $Y_{ij} = 0$ otherwise. As before, the probability of a match when alleles are drawn from the same whale is equivalent to the probability of a match when alleles are drawn from same substocks. From (2.28), if the alleles are drawn from the same whale then Y_{ij} is distributed Bernoulli$(g + (1-g)[g + (1-g)\sum_{j=1}^{m} p_j^2])$. If the alleles are drawn from different whales then Y_{ij} is distributed Bernoulli$(\sum_{j=1}^{m} p_j^2 + U[g + g(1-g)](1 - \sum_{j=1}^{m} p_j^2) + (1-U)f(1 - \sum_{j=1}^{m} p_j^2)])$. Let $Q_{ij} = P(Y_{ij} = 1)$ and

$$Z_{ij} = \log\left(\frac{Q_{ij}}{1 - Q_{ij}}\right). \tag{2.39}$$

Let us start again from the assumption of a linear relationship between f and X, where $f = \alpha_0 + \alpha_1 X_{ij}$ and X is the covariate of interest over a range of (0,1). Then I can arrange Z as

$$Z_{ij} = 2g(0) + 2(Ug(X_{ij}) - g(0))\delta_{ij} - g(0)^2 + (g(0)^2 - Ug(X_{ij})^2)\delta_{ij} + (1-U)f\delta_{ij} \tag{2.40}$$

where $\delta_{ij} = 1$ if alleles i and j are from different whales and $\delta_{ij} = 0$ otherwise. After reparameterization, (2.40) simplifies to

$$Z_{ij} = \mu + \gamma_1\delta_{ij} + \gamma_2 X_{ij}\delta_{ij} + \gamma_3 X_{ij}^2\delta_{ij}, \tag{2.41}$$

where $\mu = [2g(0) - g(0)^2]$, $\gamma_1 = (1-U)[\alpha_0 - 2\alpha_0^2 + f]$, $\gamma_2 = 2U\alpha_1(1-\theta)(1-\alpha_0)$, and $\gamma_3 = -U\alpha_1^2(1-\theta)^2$. With the same reasoning as in the independent substocks case, testing $H_o : \gamma_1 = 0$ corresponds to testing for the existence of more than one population, and testing $H_o : \gamma_2 = \gamma_3 = 0$ amounts to testing whether the covariate X affects f.

To summarize, I have developed two models for the binary random variables, Y_{ij}, for the case when there are more than two allele types. The models assert that $Y_{ij} \sim$Bernoulli(Q_{ij}) so $E(Y_{ij}) = Q_{ij}$. I have a link function,

$$Z_{ij} = \log\left(\frac{Q_{ij}}{1 - Q_{ij}}\right), \tag{2.42}$$

that does not require estimates of allele frequencies, and the linear predictor $g(Q_{ij}) = \mu + \gamma_1\delta_{ij} + \gamma_2\delta_{ij}X_{ij} + \gamma_3 X_{ij}^2\delta_{ij}$. Together these are the components of a generalized linear model in the binomial family with logit link function. Such a model can be fit easily in statistical packages such as S-Plus or R [25]. However, since the binary responses are dependent, the parameter estimates from the fitted Generalized Additive Model are not Maximum Likelihood Estimates and thus standard inferential procedures do not hold. Therefore I use permutation testing to test hypothesis about my parameters. Appropriate permutation tests of $\gamma_1 = 0$ and $\gamma_2 = \gamma_3 = 0$ can be used, respectively, to test the multiple-substock hypothesis and the hypothesis that genetic similarity depends on the covariate X.

2.3 Multi-Allele Multi-Locus Case

For analysis of multi-locus data, we limit consideration to the case where each locus can be considered as independent from the others. In other words, we assume no linkage among loci. Let there be L loci and let ℓ for $\ell = 1 \cdots L$ index loci. Then the match probabilities can be derived using the same methods as above, but with an additional index of ℓ. Furthermore, having previously shown that independent and dependent substock assumptions lead to the same generalized linear model, we treat only the independent substock case hereafter.

2.3.1 Marginal Match Probabilities for Multi-Locus Case

When there are L loci, let A_j^ℓ and p_{ij}^ℓ and m^ℓ be defined analogously to the single locus case, for $\ell = 1, \cdots, L$. Note that f is a function of the covariate, X, not of the locus. Thus, across all loci g is defined as previously, i.e. $g = f + \theta - f\theta$. Assuming loci are independent, the allele match probabilities can be found separately for each locus. That is,

$$P(A_j^\ell A_j^\ell | p_{ij}^\ell, \text{alleles from same substock}) = p_{ij}^\ell(g + (1-g)p_{ij}^\ell). \tag{2.43}$$

Let p_{ij}^ℓ be independent random variables from a Dirchlet distribution with mean vector $\mathbf{p}_j^\ell$, for $\ell = 1, \cdots, L$. If alleles drawn from different substocks are assumed to be independent then the marginal probability of a match when two alleles at locus ℓ are drawn from different substocks is

$$P(\text{match}|\text{alleles at locus } \ell \text{ from different substocks}) = \sum_{j=1}^{m^\ell} (p_j^\ell)^2. \tag{2.44}$$

If alleles drawn from different substocks are assumed not to be independent, i.e., the covariate also affects allele pair matching across substocks, then the marginal probability of a match at locus ℓ when two alleles are drawn from different substocks is

$$P(\text{match}|\text{alleles at locus } \ell \text{ from different substocks}) = f + (1-f)\sum_{j=1}^{m^\ell} (p_j^\ell)^2. \tag{2.45}$$

Let U be defined as the probability that two sampled alleles are from the same substock.

Note that U is dependent on the number of whales in each sample and does not change with respect to the number of loci and/or alleles. Thus, it will be a constant across loci. It follows that

$$P(\text{match}|\text{alleles from different whales}) = \sum_{j=1}^{m^\ell} (p_j^\ell)^2 + U[g + g(1-g)](1 - \sum_{j=1}^{m^\ell} (p_j^\ell)^2) \quad (2.46)$$

for the case when the covariate does not operate across substocks and

$$\begin{aligned} P(\text{match}|\text{alleles from different whales}) &= \sum_{j=1}^{m^\ell} (p_j^\ell)^2 + U[g + g(1-g)](1 - \sum_{j=1}^{m^\ell} (p_j^\ell)^2) \\ &+ (1-U)f(1 - \sum_{j=1}^{m^\ell} (p_j^\ell)^2) \end{aligned} \quad (2.47)$$

for the case of when the covariate operates across substocks.

2.3.2 Model for detecting stock structure from match probabilities

Again, we assume independent loci and apply the logit link function so

$$Z_{ij}^\ell = \log\left(\frac{Q_{ij}^\ell}{1 - Q_{ij}^\ell}\right), \quad (2.48)$$

where Q_{ij}^ℓ is the probability that $Y_{ij}^\ell = 1$, where Y_{ij}^ℓ is 1 if alleles drawn from locus ℓ match and zero otherwise. Thus, the structure of a generalized linear model for Z_{ij}^ℓ is suggested by these assumptions with the linear predictor

$$Z_{ij}^\ell = \mu + \gamma_0\delta_{ij} + \beta_\ell + \gamma_\ell\delta_{ij} + \alpha_0\delta_{ij}X_{ij} + \tau_0 X_{ij}^2\delta_{ij} + \alpha_\ell\delta_{ij}X_{ij} + \tau_\ell X_{ij}^2\delta_{ij} \quad (2.49)$$

where $\sum\beta_\ell = \sum\gamma_\ell = \sum\alpha_\ell = \sum\tau_\ell = 0$.

If there is no substock structure, the match probability should not be influenced by the source whale effect, δ_{ij}. Thus, the match probability for alleles drawn from same whale and from different whales should be the same. That is, testing for single substock structure is equivalent to testing the null hypothesis of $\gamma_0 = 0$. To test the significance of the covariate X_{ij}, the null hypothesis of $\alpha_0 = \tau_0 = 0$ is tested.

2.3.3 Case When Genetic Correlation Is Smooth Function of Covariate

When developing the models thus far, I have assumed that genetic correlation is a linear function of the covariate of interest. This assumption was for illustrative purposes and is not necessary. Here I will consider the most general case of genetic correlation being a smooth function of the covariate. That is, I assume that $f = \alpha_0 + \alpha_1 s(X_{ij})$ for some smooth function s, where X is the covariate of interest over a range of (0,1). Normally the covariate is defined so that $X = 0$, when alleles originate from the same source whale. Let θ and g be defined as previously, i.e., $g = \theta + f - f\theta$.

Notice that this change in the assumption about genetic correlation structure with respect to covariate X does not affect the match probabilities. Therefore, for the single-locus multi-allele case, equations (2.35)-(2.38) still hold. For the multiple-locus m-allele case equations (2.43)-(2.47) still hold. The smooth function assumption influences the nature of the linear predictor. Thus, the generalized linear model structure can be re-expressed as a generalized additive model structure, and can be fit in S-Plus or R using the 'gam' function

[55]. The linear predictor for the single-locus multi-allele case and multi-locus multi-allele case are

$$Z_{ij} = \mu + \gamma_1 \delta_{ij} + \gamma_2 \delta_{ij} s(X_{ij}) \tag{2.50}$$

and

$$Z_{ij}^{\ell} = \mu + \gamma_0 \delta_{ij} + s(X_{ij})\delta_{ij} + \beta_\ell + \gamma_\ell \delta_{ij} + s_\ell(X_{ij})\delta_{ij} \tag{2.51}$$

respectively, where $\sum \beta_\ell = \sum \gamma_\ell = \sum s_\ell = 0$. Note that, since the binary responses are dependent, the parameter estimates from fitting this model via GAM are not Maximum Likelihood Estimates and thus standard influential procedures do not hold. Therefore I use permutation testing to test hypotheses about my parameters.

2.3.4 Hypothesis Testing in Multi-Allele Multi-Locus Case

Hypothesis testing is similar to that given in the single-locus bi-allelic case. If two alleles originate from the same whale than (2.51) leads to $Z_{ij}^{\ell} = \mu + \beta_\ell$. If there exists only a single stock then the probability of two alleles from different whales at $X_{ij} = 0$ should equal that for alleles from the same whale, i.e., $\gamma_0 + s(0) = 0$. Testing for significance of the covariate X_{ij} is equivalent to testing the significance of s because $s(X)$ should be flat if the covariate has no affect on match probability.

In order to perform hypothesis testing, I use permutation tests. If the single-population hypothesis is assumed to be true then D_{ij} should have no influence on match probabilities. Therefore, shuffling of the columns in the dataset relating to D_{ij} should have no effect. The null distribution for $\gamma_0 + s(0) = 0$ can be tested by reshuffling and reanalyzing the

data. If the original estimate for $\gamma_0 + s(0) = 0$ is very unusual with respect to the null distribution then the null hypothesis is rejected and we conclude that there exists more than one population.

In order to test the significance of the covariate, I shuffle the columns in the dataset related to X. The reduced additive predictor is $Z_{ij}^{\ell} = \mu + \gamma_0\delta_{ij} + \beta_\ell + \gamma_\ell\delta_{ij} + s_\ell(X_{ij})\delta_{ij}$. If X does not influence f, then the test statistic —namely the deviance change between fitting model (2.51) and the reduced additive predictor— should be insignificant. If this test statistic is unusually different from the null distribution, then the null hypothesis is rejected and we conclude that the covariate has a significant effect in explaining the correlation structure.

In addition to deviance tests, joint coverage null confidence bands can be used to test the effect of the covariate. The pointwise confidence bands are computed from fits to the null distribution and rescaled to be joint confidence bands. If the fits go outside these bands then we conclude that the covariate effect is significant.

In the next chapter I introduce simulation methods for multi-allele multi-locus data and the results of the analysis using my method. I also show performance of my method by comparing results with competing methods.

Chapter 3

Simulation Studies

In this chapter I will study how my model, developed in Chapter 2, performs in terms of explaining population structure. This chapter is limited to simulated examples; Chapter 4 will present examples based on real data. I begin with a brief discussion of my data simulation approach and a detailed discussion of various factors related to data simulation. The bulk of this chapter describes results of models fitted to such simulated datasets. I evaluate the various simulation results and explain the contributions of using my method in terms of detecting population structure and the influence of the covariate. I also compare the results with results from other potentially applicable methods.

My simulated datasets are generated using the allele match probabilities given in Chapter 2. I simulated pairs of individuals and the corresponding covariate value. My main interest is the extent to which my method can distinguish the source(s) of correlation structure in the simulated data. When there is Hardy-Weinberg disequilibrium, my method is designed to distinguish several potential causes of it. That is, disequilibrium may be caused

by substock structure, or it may be related to a covariate, or both causes may contribute to disequilibrium. The effect of the covariate may occur in the presence or absence of multiple substocks. The importance of determining the source of correlation structure is to avoid default attribution of any and all disequilibrium to a Wahlund effect caused by multiple substocks. In the simulation studies, I assume that the influence of the covariate on within substock structure is the same in each substock for each scenario. In real applications it is possible that the covariate effect is isolated to one substock.

The data simulation starts with generating pairs of individual and assigning a covariate, X, taking values in $[0, 1]$. The covariate which influences f is generated as $X = |U_1 - U_2|$ where U_1 and U_2 are independent random variables from Uniform(0,1) distribution. I consider two different functional relationships between the covariate and f. The simplest case where f is a linear function of the covariate, i.e. $f = f(X_{ij}) = (1 - X)f_{max}$. The more biologically plausible second case allows f to be a nonlinear function of the covariate, namely $f(X) = \{1 - (10/3)^6[\max(0, X - 1/5)]^3[\min(4/5 - X, 1)]^3\}f_{max}$ where f_{max} is defined to be the strongest possible effect of the covariate on genetic correlation.

Note that θ is used to simulate multistock data and takes values in $[0, 1]$ too. If single population data are simulated, i.e; if there is no substock structure, then θ is set to zero for data simulation. When both f and θ contribute to genetic correlation, recall that the genetic correlation is parameterized by a function of the covariate, namely $g(X) = f(X)+\theta-\theta f(X)$.

The variation of allele frequencies among substocks has an important impact on our ability to detect population structure. The severity of the between-substock differences in

allele frequencies is random. Therefore, I control the strength of the signal in allele probabilities by introducing a parameter τ. Let allele probabilities for each substock be randomly sampled from a Dirchlet distribution with mean vector $\mathbf{p}^\ell$ for ℓ^{th} locus. Define ξ_τ^ℓ as the τ percentile of the distribution of S where S is the chi-squared test statistic for testing allele frequency homogeneity between two populations when cell counts exactly match the $\mathbf{p}^\ell$. Then if $\xi_{\tau+.05}^\ell \leq S \leq \xi_\tau^\ell$, the randomly drawn allele frequencies are said to have signal strength τ. Thus, large values of τ correspond to informative allele frequencies. I use τ to control signal strength when simulating allele frequencies from substocks. Quantiles like ξ_τ^ℓ are estimated via Monte Carlo.

Once substock allele frequencies are simulated, each pair of individuals can be assigned an allele pair using the allele match probabilities derived in Chapter 2. The binary variable Y which is used as the response variable in my model equals 1 if alleles match and zero otherwise.

Figure 3.1 displays the effect of τ on allele match probability. This graph shows P[match] as a function of the covariate X_{ij} when the alleles originate from different whales (solid line). The dotted line indicates the reference level of P[match] when alleles originate from the same whale. The signal represented by the Wahlund effect indicating the presence of multiple stocks is seen in these graphs as the vertical separation between the solid and dotted line. As τ decreases from left to right in Figure 3.1, we see that the slope of the solid line does not change but the distance between it and the dotted line decreases. This effect indicates greater difficulty detecting genetic signals due to stock substructure.

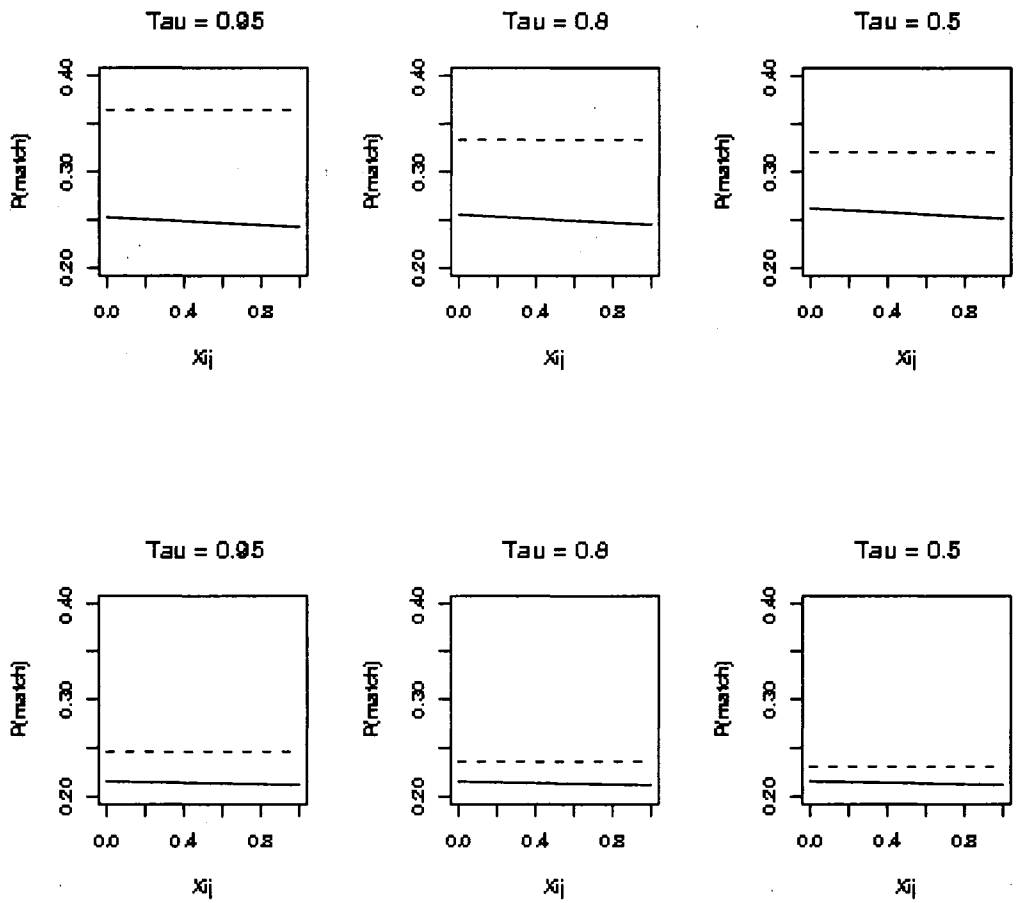

Figure 3.1: Average true P[match] (across loci and replicates) for three choices for τ when $\theta = 0.05$ and $f_{max} = 0.03$ (top row) and when $\theta = 0.01$ and $f_{max} = 0.01$ (bottom row).

3.1 Data Simulation

My method is flexible enough to be used in various cases consisting of microsatellite data. To impose some structure on simulation testing, I consider simulating data similar to the Bering-Chukchi-Beaufort Seas bowhead dataset. That is, my simulated data represent no more than two subpopulations that have a spatio-temporal migration pattern. Further, I consider difference in capture time as my covariate of interest, so allele pairs consisting of

the two alleles from the same individual will have a covariate value zero.

Each whale is simulated once and assigned a genotype based upon genotypic frequencies of the subpopulation it belongs to. Unlike with real data, multiple simulated pairings involving the same whale are avoided for simplicity (for a discussion of some implications of this choice see section 5.2.1). As discussed in the previous chapters, genetic correlation is controlled by the function $g(X) = f(X) + \theta - \theta f(X)$. Next, random alleles are chosen from each whale in the pair and they are used to determines the binary random variable Y. Y equals one if the allele pair matches and zero otherwise. I assume no linkage among loci and therefore simulate each locus independently. Accordingly, for each locus a new set of allele frequencies are generated using the overall allele frequencies and Dirchlet distribution parameter.

For allele pairs where both alleles originate from the same whale ("same-whale pairings"), X is assigned zero because the covariate is a measure of the difference of a characteristic between the whales. If the covariate was defined differently, e.g., averaging individual values, then it might take a non-zero value for the same-whale case. In such situations, the inferential strategy should be modified appropriately.

Below I introduce the different factors and scenarios that were investigated to evaluate the performance of my method when testing for population structure and for a covariate effect. The factors considered in my simulation scenarios are

- Number of substocks: 1 or 2

- τ: .95, .80 or .50
- θ : .05, .03, or .01
- f_{max}: .05, .03, or .01
- Sample sizes for single population (n): 150, 75, or 38

  Sample sizes for two population (n_1,n_2): (75,75), (38,38), or (113, 37)
- Number of loci (L): 5 or 20
- Number of alleles ($m^{(\ell)}$): 5 per locus, or Uniform$(5, 6, \cdots, 20)$ per locus
- $\mathbf{p}^\ell = (p^\ell_1, \cdots, p^\ell_{m^\ell})$: $(\frac{1}{m^\ell}, \cdots, \frac{1}{m^\ell})$ or Dirchlet$(\mathbf{1})$.

The baseline two population scenario consists of $\theta = .05$, $f_{max} = .03$, (n_1,n_2)=(75,75), 5 loci with 5 allele per locus, $\mathbf{p}^\ell = (1/5, \cdots, 1/5)$ and $\tau = .95$. The baseline for single population scenario is the same as that of two population but consists of $\theta = 0$ and $n = 150$ instead of $\theta = .05$ and (n_1,n_2)=(75,75). Each scenario is repeated 10 times, and 500 permutations are used at each replicate to calculate the p-values. These choices were made to keep the total computing time manageable.

I have created some figures to help visualize the variation in match probabilities in the simulation scenarios I have set up. Figure 3.2 shows the true match probabilities for each locus and also averaged over the 5 loci, for the case when $\theta = .05$, $f_{max} = .03$, and remaining parameters are at baseline values. The dotted line represents the true match probability when both alleles are from the same whale. This actually is a point on the plot since $X = 0$ for this case and so the dotted line should be considered as a reference line. The solid line represents the true match probability when allele pairs are from different

whales. Evaluating the plots, it can be concluded that the variation among loci can be substantial. Next, Figure 3.3 shows the true match probability fit for the 10 replicates averaged over the 5 loci and the match probability plot averaged over both 10 replicates and 5 loci. Here, it can be concluded that between replicate variation in locus-averaged match probabilities is quite low.

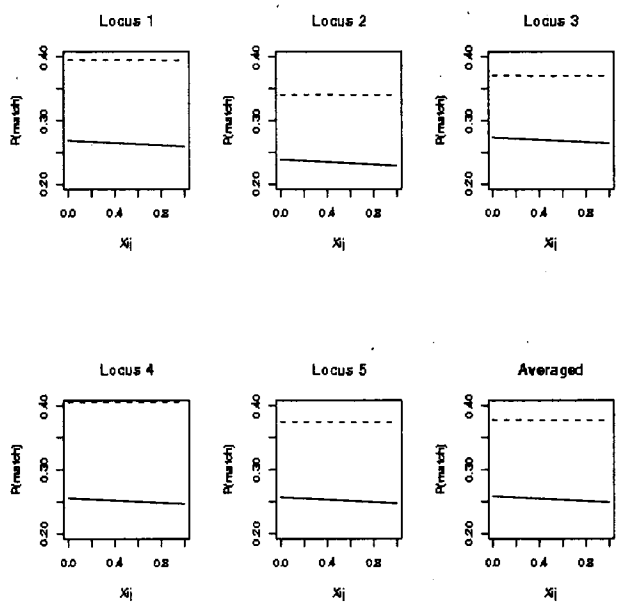

Figure 3.2: True P[match] for each locus for the scenario with $\theta = 0.05$, $f_{max} = 0.03$, and remaining parameters at baseline values. The final panel shows the average across loci. Dotted lines indicate P[match] when both alleles are from the same individual, and solid lines show P[match] when the alleles originate from different individuals.

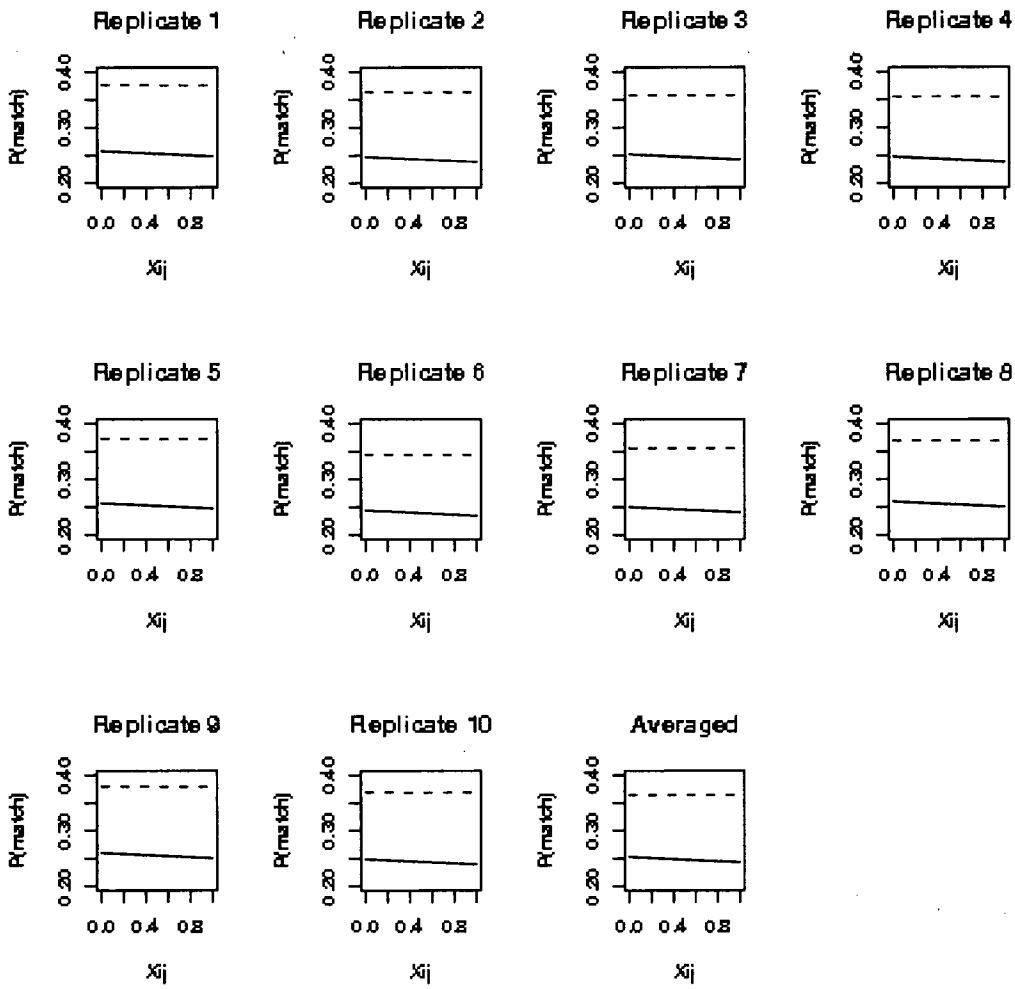

Figure 3.3: Average true P[match] (across loci) for each of 10 replicates of the scenario used in Figure 3.2, along with the overall average over replicates. Dotted lines indicate P[match] when both alleles are from the same individual,and solid lines show P[match] when the alleles originate from different individuals.

As a basic check to confirm that my method provides reasonable estimates of the true match probabilities, I examined Figure 3.4, which shows the fitted match probability plot. These fits appear to be good estimates of the true match probabilities shown in Figure 3.2. It is worth emphasizing that sampling variability in the estimate of the dotted line (i.e.

same-whale match probability) is much greater than for the main portion of the model (solid line). This is because the sample size for same-whale allele pairings is several orders of magnitude smaller than for different-whale pairings. Thus, the reality-check provided by Figure 3.4 pertains mainly to the solid line.

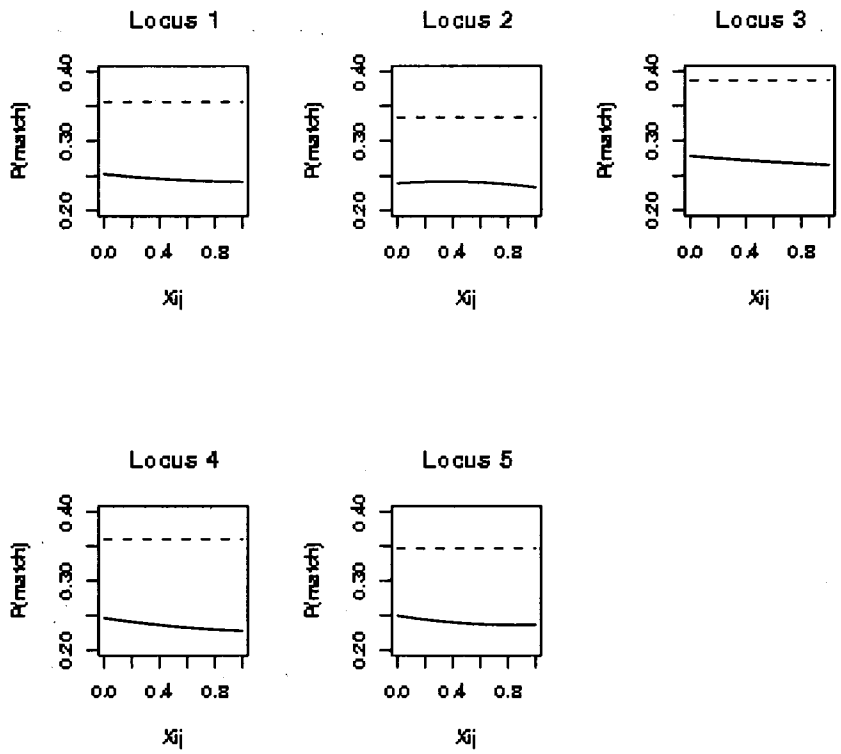

Figure 3.4: Fitted P[match] for each locus for the scenario with $\theta = 0.05$, $f_{max} = 0.03$, and remaining parameters at baseline values. Dotted lines indicate P[match] when both alleles are from the same individual, and solid lines show P[match] when the alleles originate from different individuals.

To visualize the range of covariate signals captured by my simulation scenarios, I created Figure 3.5 (which focuses on two-stock scenarios). The same plots are given for one-stock scenarios in Figure 3.6. For the single population scenario, the true match probability at

$X = 0$ and that for same whale is the same. Thus, the dotted line and the solid line intercept at the same point on the true match probability plots. In both single- and two-population scenarios, the slope of the solid line determines the strength of the covariate effect. It should be noted that allele pairs from different whales influence the slope and allele pairs from same whales influence the intercept. The number of allele pairs from different whales is much more than the number from same whales, resulting in more power for detecting significance of covariate than for detecting significance of population structure through D_{ij}. Since $\theta = 0$ for the case of no substock structure, the panels in Figure 3.6 represent cases of increasing f_{max} values (towards the right), with θ kept fixed at zero. An increase in f_{max} is associated with an increase in the magnitude of the slope of the different whale match probability line. This figure shows that the one-population scenarios include cases where it will be very difficult to detect the effect of the covariate.

In Figure 3.5, θ increases towards the right. This results in a larger gap at the intersection of the two match probability lines. f_{max} increases towards the bottom resulting in steeper slope of the (solid) different whales match probability line. This figure shows that the range of scenarios I examined includes many very challenging scenarios where the signals my model is trying to detect are very small.

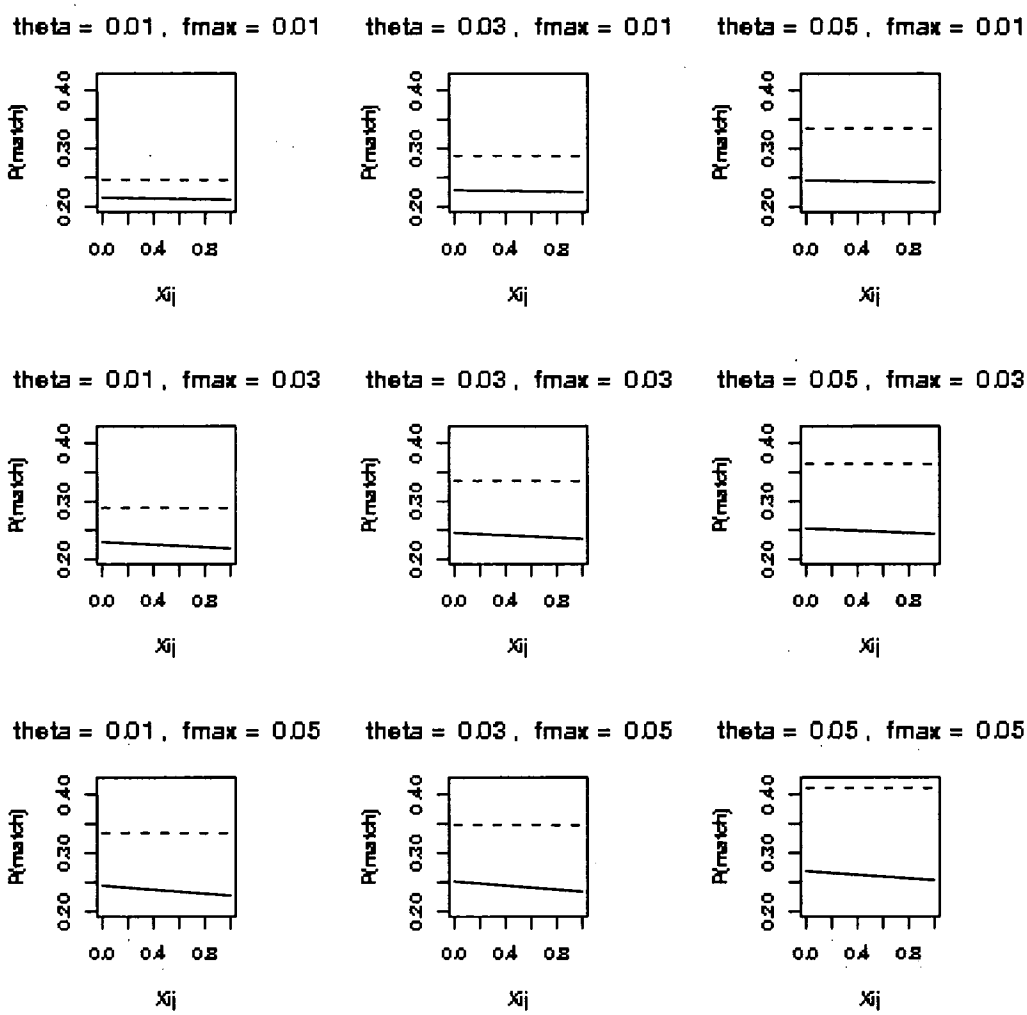

Figure 3.5: True P[match] for two-population linear dependence with θ increasing from left to right,i.e. $\theta = 0.01, 0.03$ and 0.05, and f_{max} increasing from top to bottom, i.e., $f_{max} = 0.01, 0.03$ and 0.05, averaged over the five loci in each scenario. The effect of increasing θ is seen by comparing plots in the same row. The effect of increasing f_{max} is seen by comparing plots in the same column.

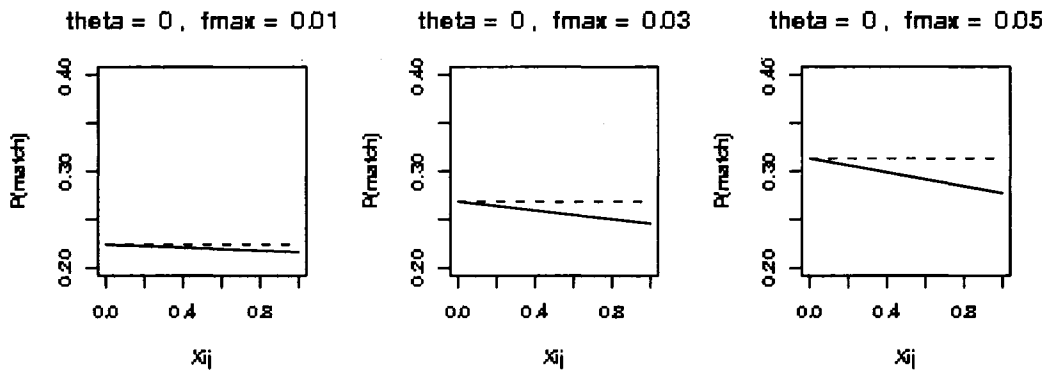

Figure 3.6: True P[match] for one-population linear dependence with $\theta = 0$, $f_{max} = 0.01, 0.03$ and 0.05 averaged over the five loci in each scenario. The strength of the covariate effect is related to the slope of the solid line.

3.2 Case 1: Single-population with Linear Dependence of f on X

My first simulation study examines the case when there is only a single population and the dependence of f on X is linear according to $f(X) = (1 - X)f_{max}$. For such datasets, the model given in equation (2.49) was fit. Results are summarized in Table 3.1. This table shows the median p-values for testing for an effect of X (the covariate) and D (indicating multiple subpopulations) over the 10 replicate runs of the indicated scenarios. Additionally, counts of the numbers of significant p-values (at the 0.05 level) among the 10 replicates are also given. It can be seen that my method usually detects significant covariate effects and gives almost no indication of population substructure for the various scenarios. When the effect of the covariate is quite small ($f_{max} = .01$), Table 3.1 shows that the power to detect the effect of the covariate is somewhat diminished. The smaller sample size ($n_1 = 38$)

also show diminished power for detecting the covariate effect. When the number of loci is increased to 20, the power of my method is considerably higher with $p < 0.001$ for every replicate. Overall, my method performs quite well for the various changing parameters such as θ, sample size and f_{max}, and almost never commits a Type I error by falsely detecting multiple stocks.

To check how well my model estimates the true match probabilities, I consider comparing the true match probability plot given in Figure 3.7 and the fitted match probability plot given in Figure 3.8. As mentioned previously, the dotted line represents the same-whale match probability while the solid line is that for different-whale pairings. The small number of same-whale pairings with respect to the number of different whale pairings explains the larger variation in the same-whale match probabilities from one locus to the other. This comparison again confirms that my code is correctly implementing the intended model and estimation procedure.

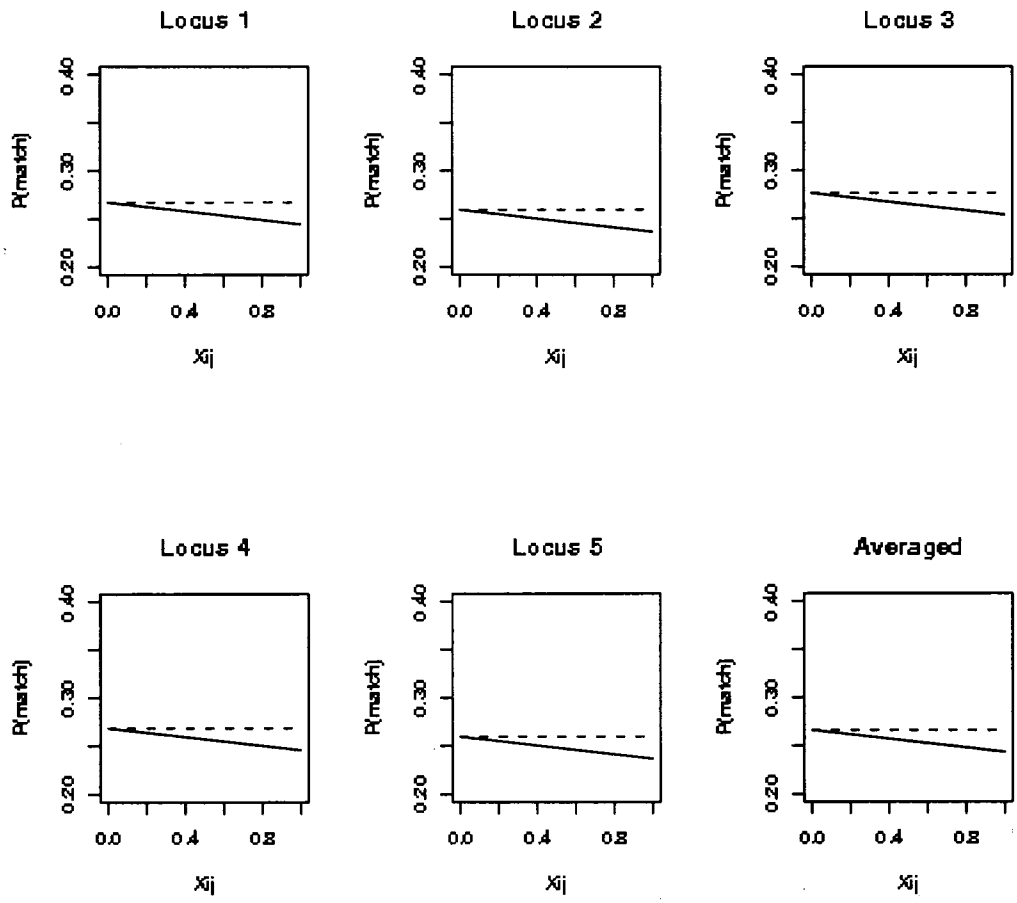

Figure 3.7: True P[match] for each locus for the single-population scenario with $\theta = 0$, $f_{max} = 0.03$, and remaining parameters at baseline values. Dotted lines indicate P[match] when both alleles are from the same individual, and solid lines show P[match] when the alleles originate from different individuals.

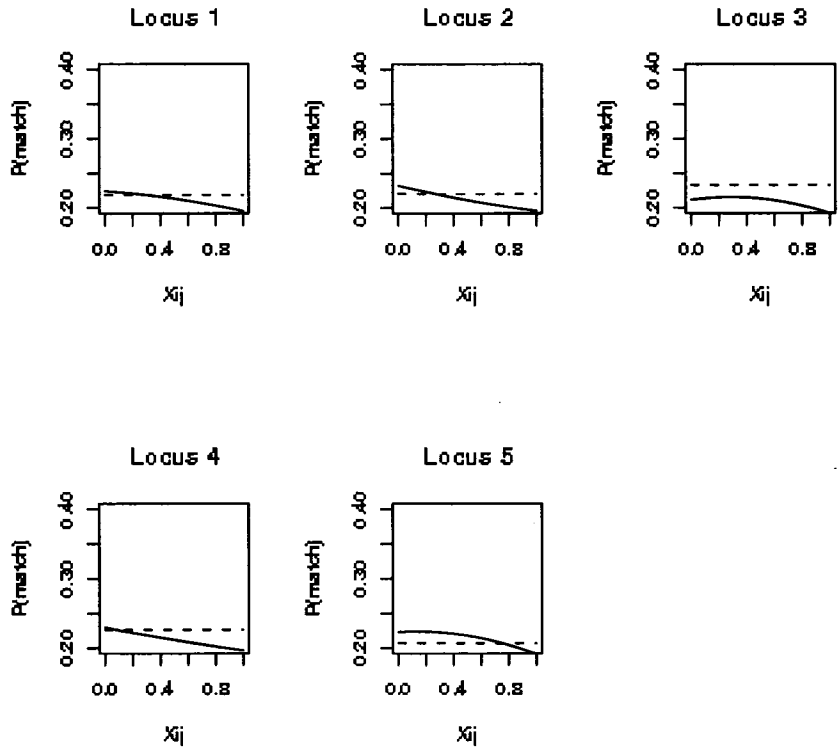

Figure 3.8: Fitted P[match] for each locus for the single-population scenario with $\theta = 0$, $f_{max} = 0.03$, and remaining parameters at baseline values. Dotted lines indicate P[match] when both alleles are from the same individual, and solid lines show P[match] when the alleles originate from different individuals.

Table 3.1: One-population simulation results for fitting (2.49) to data generated with a linear dependence between f and the covariate. Scenarios are specified by their deviations from the baseline. The number of significances is the number of times out of the ten replicates where the p-value did not exceed 0.05.

Scenario specification	Median p-value		Number of significances	
	for X_{ij}	for D_{ij}	for X_{ij}	for D_{ij}
Baseline	0.000	0.404	10	1
$\tau = 0.8$	0.000	0.593	10	0
$\tau = 0.5$	0.000	0.385	10	0
$f_{max} = 0.05$	0.000	0.591	10	0
$f_{max} = 0.05$, $\tau = 0.8$	0.000	0.463	10	0
$f_{max} = 0.05$, $\tau = 0.5$	0.000	0.716	10	0
$f_{max} = 0.01$	0.009	0.259	8	0
$f_{max} = 0.01$, $\tau = 0.8$	0.022	0.503	7	0
$f_{max} = 0.01$, $\tau = 0.5$	0.011	0.614	7	0
$n_1 = 38$	0.027	0.465	7	0
$n_1 = 38$, $\tau = 0.8$	0.013	0.582	5	1
$n_1 = 38$, $\tau = 0.5$	0.040	0.390	5	0
$L = 20$	0.000	0.408	10	0
$L = 20$, $\tau = 0.8$	0.000	0.825	10	0
$L = 20$, $\tau = 0.5$	0.000	0.748	10	0

3.3 Case 2: Two-subpopulations with Linear Dependence of f on X

My second simulation study examines the case when there are two genetically distinct subpopulations and the dependence of f on X is linear, as above. Some results for such scenarios are given in Table 3.2. Generally, my method is most successful in detecting the covariate effect for scenarios with $f_{max} > .01$ and in detecting population structure with $\theta > .01$ and $\tau > .5$.

In order to further examine the influence of various factors, additional scenarios were examined. These are summarized in Table 3.3, which breaks down the scenarios into three groups. The first group of scenarios in Table 3.3 varies the number of alleles per loci and lets allele frequencies vary among loci. This is more close to real data situations. The baseline sample sizes, $(n_1, n_2) = (75, 75)$, vary in the second group of scenarios in Table 3.3. Finally, the bottom section of Table 3.3 examines the power of detecting very small effects by changing values for θ, f_{max}, number of loci, and allele frequencies.

Comparing the results in Table 3.2 to the results in Table 3.3, it can be seen that the increase in the number of loci from 5 to 20 increases the detected significance of X_{ij}. Moreover, when the number of alleles per loci are random, the significance of X_{ij} increases, while a minor decrease in the significance of D_{ij} is observed. Another comparative result relates to the influence of sample size. There is an obvious decrease in the detected significance of both X_{ij} and D_{ij} when sample size is decrease from $(n_1, n_2) = (75, 75)$ to $(n_1, n_2) = (38, 38)$.

Table 3.2: Two-population simulation results for fitting (2.49) to data generated with a linear dependence between f and the covariate. All parameters except θ, f_{max}, and τ were kept at baseline values. The number of significances is the number of times out of the ten replicates where the p-value did not exceed 0.05.

Scenario specification			Median p-value		Number of significances	
f_{max}	θ	τ	for X_{ij}	for D_{ij}	for X_{ij}	for D_{ij}
0.05	0.05	0.95	0.000	0.000	10	10
0.05	0.03	0.95	0.000	0.000	9	10
0.05	0.01	0.95	0.000	0.000	10	10
0.03	0.05	0.95	0.005	0.000	9	10
0.03	0.03	0.95	0.000	0.000	9	9
0.03	0.01	0.95	0.000	0.000	10	9
0.01	0.05	0.95	0.375	0.000	1	10
0.01	0.03	0.95	0.471	0.000	1	10
0.01	0.01	0.95	0.232	0.196	1	2
0.05	0.05	0.80	0.000	0.000	10	10
0.05	0.03	0.80	0.000	0.001	10	10
0.05	0.01	0.80	0.000	0.019	10	6
0.03	0.05	0.80	0.010	0.000	7	10
0.03	0.03	0.80	0.007	0.003	9	9
0.03	0.01	0.80	0.006	0.092	9	2
0.01	0.05	0.80	0.406	0.012	1	9
0.01	0.03	0.80	0.268	0.015	1	9
0.01	0.01	0.80	0.147	0.157	1	2
0.05	0.05	0.50	0.000	0.000	10	10
0.05	0.03	0.50	0.000	0.010	10	6
0.05	0.01	0.50	0.000	0.060	10	5
0.03	0.05	0.50	0.006	0.004	8	10
0.03	0.03	0.50	0.001	0.024	10	6
0.03	0.01	0.50	0.000	0.628	10	2
0.01	0.05	0.50	0.259	0.039	0	5
0.01	0.03	0.50	0.340	0.309	1	1
0.01	0.01	0.50	0.240	0.510	3	0

Table 3.3: Two-population simulation results for fitting (2.49) to data generated with a linear dependence between f and the covariate. Scenarios are specified by their deviations from the baseline. The number of significances is the number of times out of the ten replicates where the p-value did not exceed 0.05. The first section explores more realistic locus characteristics; the second section explores sample size; the third section explores power to detect tiny effects.

Scenario specification	Median p-value		Number of significances	
	for X_{ij}	for D_{ij}	for X_{ij}	for D_{ij}
$m^{(\ell)}$ random	0.000	0.000	10	10
$m^{(\ell)}$ random, $\tau = 0.8$	0.000	0.000	10	9
$m^{(\ell)}$ random, $\tau = 0.5$	0.000	0.000	10	9
$\mathbf{p}^{(\ell)}$ random	0.008	0.000	8	10
$\mathbf{p}^{(\ell)}$ random, $\tau = 0.8$	0.030	0.001	6	10
$\mathbf{p}^{(\ell)}$ random, $\tau = 0.5$	0.028	0.015	6	7
$L = 20$	0.000	0.000	10	10
$L = 20$, $\tau = 0.8$	0.000	0.000	10	10
$L = 20$, $\tau = 0.5$	0.000	0.000	10	10
$n_1 = n_2 = 38$	0.142	0.000	1	10
$n_1 = n_2 = 38$, $\tau = 0.8$	0.150	0.024	4	6
$n_1 = n_2 = 38$, $\tau = 0.5$	0.128	0.063	1	3
$n_1 = 37$, $n_2 = 113$	0.000	0.000	9	10
$n_1 = 37$, $n_2 = 113$, $\tau = 0.8$	0.001	0.007	9	7
$n_1 = 37$, $n_2 = 113$, $\tau = 0.5$	0.000	0.030	10	6
$n_1 = n_2 = 38$, $L = 20$, $m^{(\ell)}$ random	0.000	0.000	10	10
$n_1 = n_2 = 38$, $L = 20$, $m^{(\ell)}$ random, $\tau = 0.8$	0.000	0.001	10	10
$n_1 = n_2 = 38$, $L = 20$, $m^{(\ell)}$ random, $\tau = 0.5$	0.000	0.001	10	9
$\theta = 0.01$, $f_{max} = 0.01$, $L = 20$	0.033	0.066	7	3
$\theta = 0.01$, $f_{max} = 0.01$, $L = 20$, $\tau = 0.8$	0.013	0.194	7	3
$\theta = 0.01$, $f_{max} = 0.01$, $L = 20$, $\tau = 0.5$	0.023	0.336	8	0
$\theta = 0.01$, $f_{max} = 0.01$, $m^{(\ell)}$ random	0.045	0.285	5	2
$\theta = 0.01$, $f_{max} = 0.01$, $m^{(\ell)}$ random, $\tau = 0.8$	0.067	0.170	4	4
$\theta = 0.01$, $f_{max} = 0.01$, $m^{(\ell)}$ random, $\tau = 0.5$	0.100	0.241	2	2

This decrease in the power due to smaller sample size is negated when number of loci is increased from 5 to 20. When the sample size is changed to $(n_1, n_2) = (37, 113)$ the significance of D_{ij} slightly decreases while that for X_{ij} slightly increases. Thus it can be said that power of detecting population structure and the covariate effect is influenced by number of loci, number of allele per loci, and sample size.

3.4 Cases 3 and 4: One and Two Subpopulations with Nonlinear Dependence of f on X

In the previous section, I simulated data such that there was a linear relationship between f and the covariate, X. My analysis method is flexible in that it is not restricted to such a linear relationship. In this section I generate data such that $f = f(X) = \{1 - (10/3)^6[\max(0, X - 1/5)]^3[\min(4/5 - X, 1)]^3\} f_{max}$ where f_{max} is as defined before. Figure 3.9 shows a plot of this non-linear function for choices of f_{max} equal to 0.01, 0.03 and 0.05. The shape of this function was specifically designed to mimic the findings of [42] in some bowhead whale data, and to provide a challenging and biologically plausible test of my approach.

As for this nonlinearity assumption, my baseline parameter values remain as described in previous sections. Data simulation steps are also the same except that the linear function is replaced by the nonlinear function of the covariate. Previously, p-values were calculated using deviance tests. Here, I use both deviance tests and 95% joint confidence bands to

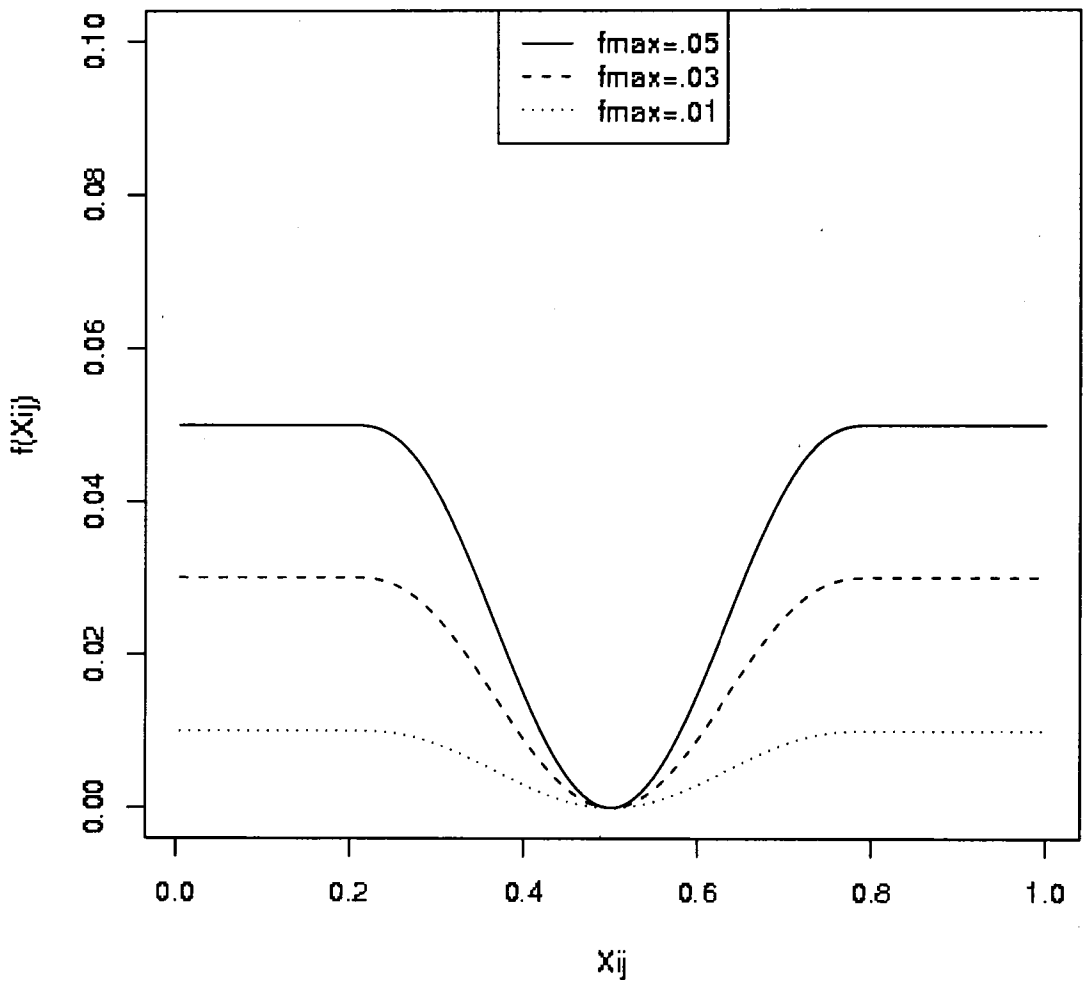

Figure 3.9: Plot of the nonlinear function $f(X_{ij}) = \{1 - (10/3)^6[\max(0, X_{ij} - 1/5)]^3[\min(4/5 - X_{ij}, 1)]^3\}f_{max}$ where f_{max} takes values $0.05, 0.03$ or 0.01.

calculate p-values, as discussed in Section 2.3.4.

To visualize the range of covariate signals captured by my simulation scenarios when the linear function is replaced by the nonlinear function of the covariate, I created Figure 3.10. f_{max} increases towards the bottom resulting in an increase in the magnitude of the hump seen in the different-whale match probability line. Similar to the case in Figure 3.5, θ in-

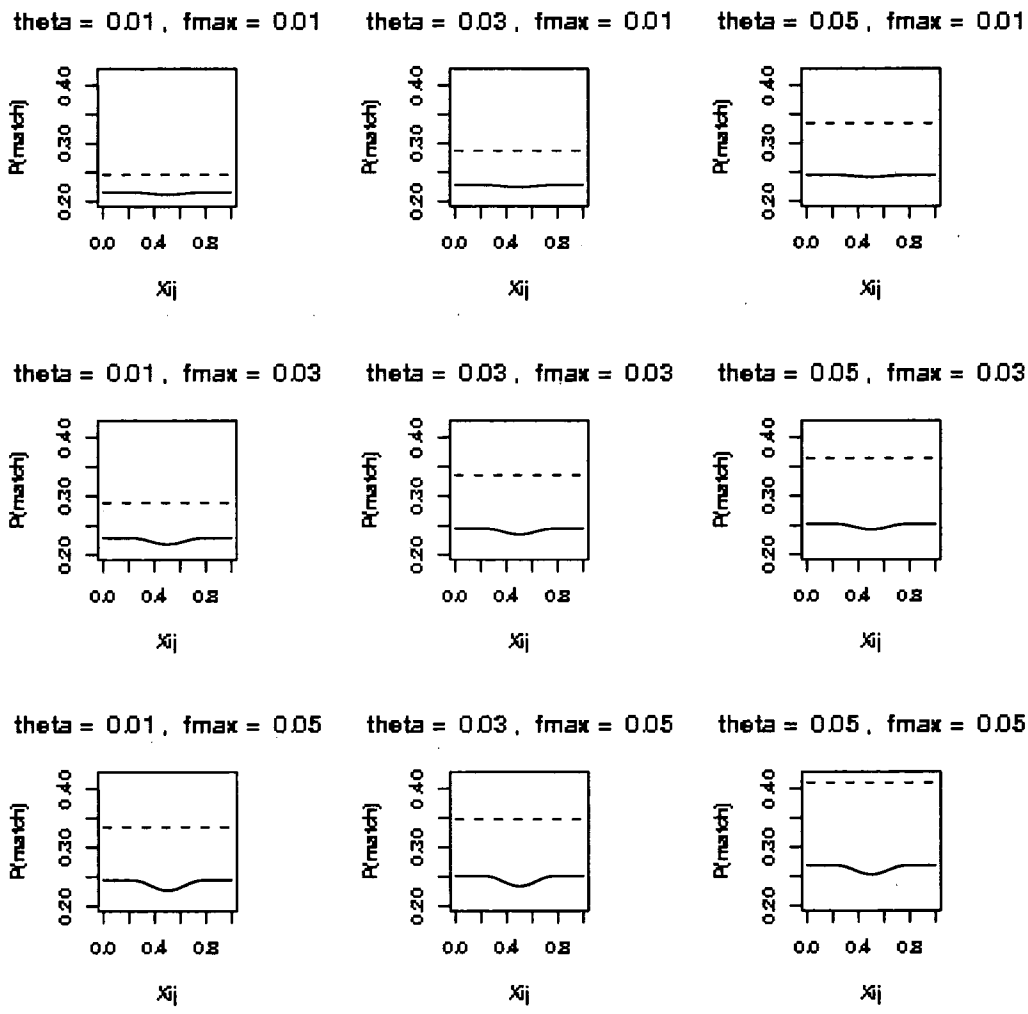

Figure 3.10: Average true P[match] (across loci and replicates) for six scenarios described in the text.

creases towards the right resulting in larger gaps between the two match probability lines. The range of scenarios I examined when f_{max} is associated with a non-linear function of the covariate include many very challenging scenarios where the signals my model is trying to detect are very small.

The results for baseline and different scenario runs are given in Table 3.4. The fully

Table 3.4: One- and two-population simulation results for fitting (2.51) to data generated with a nonlinear dependence between f and the covariate. The one-population runs are in the top portion of the table. Scenarios are specified by their deviations from the baseline. The number of significances is the number of times out of the ten replicates where the p-value did not exceed 0.05. For p-values, the column labeled 'bands' refer to assessing significance from the joint confidence bands rather than the deviance test. For ILS, 'any' refers instances in which the effect of X_{ij} was significant using either the deviance test or the 'bands' threshold.

Scenario specification	Median p-value			Number of significances		
	for X_{ij}	bands	for D_{ij}	for X_{ij}	any	for D_{ij}
Baseline, one pop.	0.000	0.000	0.515	10	10	0
$f_{max} = 0.05$	0.000	0.000	0.584	10	10	0
$f_{max} = 0.01$	0.004	0.005	0.478	8	7	0
$n_1 = 38$	0.037	0.040	0.527	6	6	1
$n_1 = 75$	0.000	0.002	0.519	10	10	1
Baseline, two subpops	0.011	0.006	0.000	8	8	10
$\theta = 0.05$, $f_{max} = 0.05$	0.000	0.000	0.000	10	10	10
$\theta = 0.01$, $f_{max} = 0.01$	0.477	0.476	0.036	2	1	6
$\mathbf{p}^{(\ell)}$ random	0.147	0.128	0.000	3	3	10
$m^{(\ell)}$ random	0.000	0.000	0.002	10	10	10
$n_1 = n_2 = 38$	0.156	0.121	0.001	0	1	9
$\theta = 0.05$, $f_{max} = 0.05$, $n_1 = n_2 = 38$	0.058	0.098	0.000	3	3	10

general model given in (2.51) was fit. In the table, the column labeled 'bands' represents the p-value from 95% joint confidence bands and 'any' represents the number of significant p-values using both deviance tests and 95% joint confidence bands. The top section of Table 3.4 shows results for single-population scenarios and the bottom section has results for two-population runs.

The baseline runs successfully detect a significant covariate effect, in both single- and two-population cases. Also for the baseline cases the results for population structure are consistent with the underlying true stock structure, in that two stocks are detected in the two-population scenario but not in the one-population scenario.

Comparing the baseline runs and other scenario runs, the influence of sample size, $\mathbf{p}^{(\ell)}$ and different numbers of allele per locus can be seen. From the top portion of Table 3.4 it can be seen that for the single population scenarios, decreasing sample size from $n_1 = 150$ to 75 has a small effect on the significance of the covariate effect and decreasing sample size further has even more impact on detection of the covariate effect. For the two population scenario, when sample size is decreased from $n_1 = n_2 = 75$ to $n_1 = n_2 = 38$, the covariate effect is no longer significant. A similar effect can be seen when both θ and f_{max} is decreased to 0.01 for the two population scenario, resulting in decreasing power of detecting significant covariate effect and D_{ij}. This is expected since $\theta = f_{max} = 0.01$ is a challenging scenario for my model to detect.

In order to see how well my model fits the true match probabilities I plotted the baseline scenario and the challenging scenario of $\theta = f_{max} = 0.01$. Figure 3.11 shows the true match

probability plot for these two scenarios. The corresponding fitted match probability plots are given in Figure 3.12. These plots show how effective my model fitting approach can be for detecting the general shape of the covariate dependence, and also illustrates again the extreme challenge of cases where $f_{max} = 0.01$

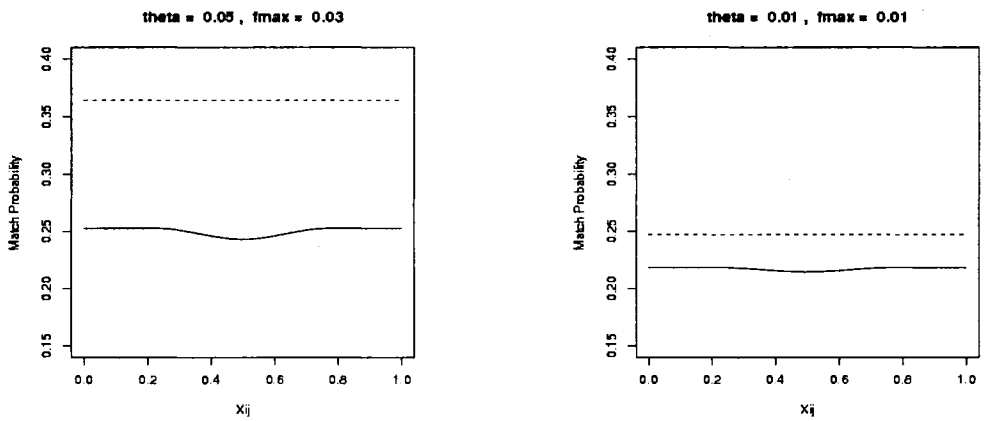

Figure 3.11: Average true P[match] (across loci and replicates) for two scenarios with $\theta = 0.05, f_{max} = 0.03$ on the left side plot and $\theta = 0.01, f_{max} = 0.01$ on the right side plot having nonlinear dependence between the covariate and P[match].

Figure 3.13 shows the fitted match probability plots for the two-population baseline scenario when f_{max} is associated with a non-linear function of the covariate. The plots consist of nine replications of the baseline scenario. In all nine plots, there is an obvious hump in the different whale match probability (solid) line. As can be seen from Table 3.4, the covariate is not significant in only two of the 10 replications, one of which is plotted in Figure 3.13 with the solid line staying within the 95% confidence band. Moreover, the gap of the two match probability lines are large enough to detect population structure as supported by results in Table 3.4.

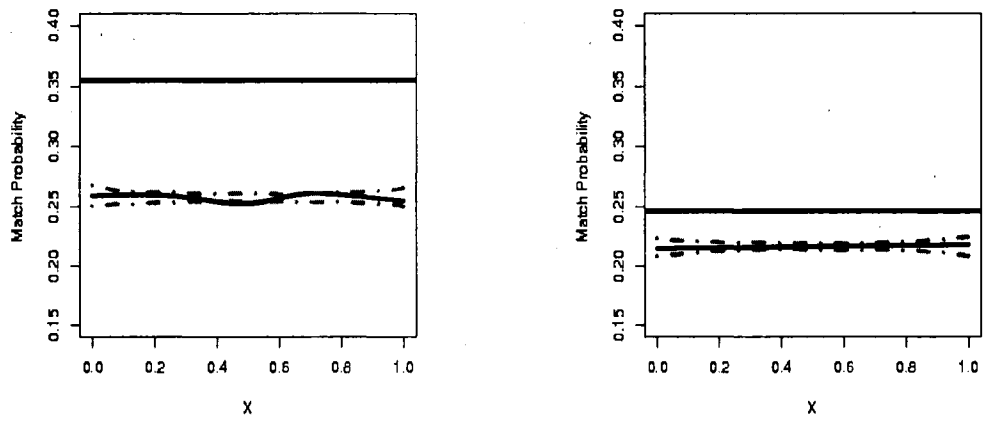

Figure 3.12: Fitted P[match] for two-population nonlinear dependence with $\theta = 0.05, f_{max} = 0.03$ on the left side plot and $\theta = 0.01, f_{max} = 0.01$ on the right side plot. The dotted curves represent estimated 95% joint confidence bands for the fit.

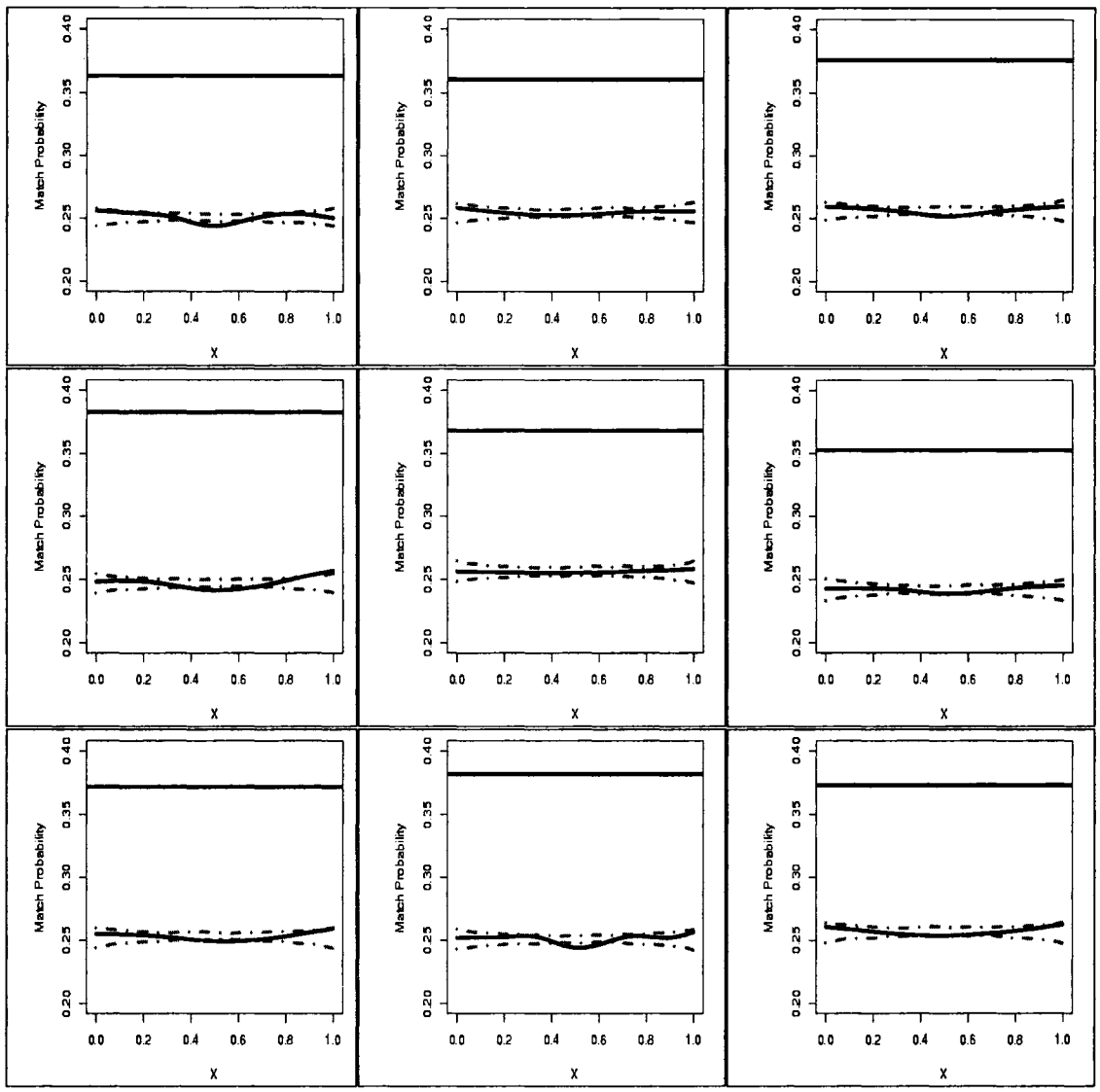

Figure 3.13: Nine of the 10 fitted P[match] plots for two-population non-linear dependence baseline runs with $\theta = 0.05$, $f_{max} = 0.03$ and $\tau = .95$.

3.5 General conclusions from simulation studies

From the results presented above, it can be concluded that my method is generally effective in detecting the genetic correlation structure and distinguishing the influence of the covariate and the Wahlund effect. As with any statistical method, power is greater for detecting strong signals than for weak signals. Strong signals correspond to large effect sizes, large sample sizes, large numbers of variable loci, and so forth.

For all three major cases (namely, single-population with linear dependence, two-population with linear dependence, and two-population with non-linear dependence), sample size has an important impact in detecting the significance of covariate effect and/or Wahlund effect. The loss in power due to decreased sample size can be balanced by an increase in the number of loci. For example, the results in Tables 3.2 and Table 3.3 for $f_{max} = 0.01$ and $\theta = 0.01$ show virtually complete failure to detect the covariate effect with $L = 5$ but about 70% power to detect it when $L = 20$. Random number of alleles per loci also seems to increase the power of detecting covariate effect. Uneven allele frequencies (i.e. large values of $||f^{\ell} - \frac{1}{m^{\ell}}\mathbf{1}||)\ \mathbf{p}^{(\ell)}$ have a tendency to decrease the power. Neither random number of alleles per loci or uneven allele frequencies has a major impact on the power of detecting the Wahlund effect.

As a result it can be concluded that my method is successful in distinguishing the sources of correlation structure under various scenarios such as changing sample size, number of loci, number of alleles per loci or random $\mathbf{p}^{(\ell)}$. It is recommended that the number of

loci be kept as large as possible, especially when the number of individuals data is collected from is small.

3.6 Comparison with Related Methods

3.6.1 Tests for disequilibrium

Here, I compare the performance of my method with several other methods for the analysis of microsatellite data. I use the same data analyzed above in the baseline scenarios, obtained by saving 150 same-whale random allele pairs of the simulated baseline scenarios. Since each baseline scenario was replicated 10 times above, this provides 10 replicate datasets for analysis here.

My first approach is to search for Hardy-Weinberg disequilibrium using the Monte Carlo testing approach implemented in the GENEPOP package [47, 48]. These MCMC analysis were run with chain lengths of 1 million, 1000 batches and 30,000 burn in. I obtained p-values for each of the 10 replicates. Table 3.5 shows the p-values for one of the 10 replicates for the three baseline runs. The heterozygote deficiency p-value is a one-sided p-value where the alternative hypothesis is that there is a deficiency of heterozygotes. On the other hand, the disequilibrium p-value is a two sided p-value where the alternative hypothesis is that Hardy-Weinberg equilibrium does not hold due to deficiency or excess of heterozygotes. For each of the three baseline scenarios, the p-values are given for each locus, followed by an overall p-value in Table 3.5. The overall p-value is obtained via Fisher's method [46]

The results shown in Table 3.5 show significant Hardy-Weinberg disequilibrium for the two-population cases but not for the one-population case. Moreover, the two-population signal is easily detected at each separate locus. Therefore, it suffices hereafter to examine only the overall p-values that are pooled across loci. This will permit us to examine more easily performance across replicates.

Table 3.6 gives the overall p-values when each of the 10 single population baseline replicates are analyzed. GENEPOP indicates significance 4 out of the 10 times. If positive outcomes are interpreted as departure from Hardy-Weinberg equilibrium due to Wahlund effect, then Genepop made four Type I errors. If significant outcomes are interpreted as indicating a different source of heterozygote deficiency, then Genepop made 6 errors due to lack of detecting this effect. Furthermore, this approach is not able to link any findings to the effect of the covariate. In contrast, my method made only 1 Type I error detecting a Wahlund effect. Furthermore, my approach is able to link the heterozygote deficiency to the covariate on all 10 of the 10 replicate runs.

Table 3.7 and Table 3.8 show similar results for the two-population linear and non-linear baseline cases, respectively. The GENEPOP p-values were significant and thus indicated major departures from Hardy-Weinberg equilibrium. This is correct but is lacking any additional information about the effect of the covariate. My method breaks down the source of departure from Hardy-Weinberg equilibrium into a portion due to the influence of a covariate and a portion due to Wahlund effect after controlling for the influence of the covariate.

Table 3.5: P-values for Hardy-Weinberg equilibrium tests for the baseline scenarios of one- and two-populations with linear and non-linear dependence between f and the covariate.

One-Population		
Locus	Heterozygote Deficiency	Disequilibrium
locus 1	0.3023	0.6096
locus 2	0.1179	0.6364
locus 3	0.5034	0.5111
locus 4	0.3623	0.3779
locus 5	0.1943	0.3714
All loci	0.6566	0.5084
Two-population with linear dependence		
locus 1	<.0001	0.0002
locus 2	0.0003	<.0001
locus 3	0.0001	<.0001
locus 4	<.0001	<.0001
locus 5	<.0001	<.0001
All loci	<.0001	<.0001
Two-population with nonlinear dependence		
locus 1	0.0002	<.0001
locus 2	<.0001	<.0001
locus 3	0.0001	<.0001
locus 4	0.0011	<.0001
locus 5	0.0004	<.0001
All loci	<.0001	<.0001

3.6.2 Permutation χ^2 tests for allele frequency differences between strata

Another common tool for the analysis of population structure using microsatellite data is permutation χ^2 comparison of allele frequencies between strata. Such methods are implemented, for example, in GENEPOP [47, 48]. Here, I examine this approach.

In practice the population identity is unknown and therefore a priori assignment to strata for permutation χ^2 analysis is largely a matter of guesswork unless assignment hypotheses can be generated from clear biological evidence or reasoning. Misclassification is likely, and misassignment of individuals to strata can influence inference about the population structure. I will consider the effect of such misclassification as part of my comparative Genepop analysis. Since the analysis method I propose in this dissertation does not rely on population identity, such misclassification does not have any influence on the resulting p-values.

For the GENEPOP runs, I consider three misclassification cases; (1) all allele pairs are stratified correctly, (2) two-thirds of each substock is stratified correctly with the remaining animals mistakenly placed in the wrong stratum, and (3) half of each substock is classified correctly while the other half is misclassified. The resulting p-values for each case are given in Tables 3.9- 3.11. In all of the tables, p-values for my method are given

Table 3.6: P-values for Hardy-Weinberg equilibrium tests using the one-population baseline datasets with linear dependence between f and the covariate. All 10 replicates are shown.

	My method p-value		Genepop p-values	
	for X_{ij}	for D_{ij}	for Heterz. Def.	for Prob. Test
Rep 1	<.0001	0.532	0.1579	0.5808
Rep 2	<.0001	0.400	0.0111	0.4359
Rep 3	<.0001	0.858	0.2478	0.7666
Rep 4	<.0001	0.408	0.0185	0.6354
Rep 5	<.0001	0.004	0.1207	0.0873
Rep 6	<.0001	0.294	0.0998	0.7058
Rep 7	<.0001	0.984	0.1749	0.4100
Rep 8	<.0001	0.700	0.1219	0.0063
Rep 9	<.0001	0.378	0.0142	0.0122
Rep 10	<.0001	0.310	0.0305	0.6347

Table 3.7: P-values for Hardy-Weinberg equilibrium tests using the two-population baseline datasets with linear dependence between f and the covariate. All 10 replicates are shown.

	My method p-value		Genepop p-values	
	for X_{ij}	for D_{ij}	for Heterz. Def.	for Prob. Test
Rep 1	0.018	<.0001	<.0001	<.0001
Rep 2	<.0001	<.0001	<.0001	<.0001
Rep 3	0.006	<.0001	<.0001	<.0001
Rep 4	<.0001	<.0001	<.0001	<.0001
Rep 5	0.024	<.0001	<.0001	<.0001
Rep 6	0.004	<.0001	<.0001	<.0001
Rep 7	<.0001	<.0001	<.0001	<.0001
Rep 8	0.054	<.0001	<.0001	<.0001
Rep 9	0.012	<.0001	<.0001	<.0001
Rep 10	<.0001	<.0001	<.0001	<.0001

Table 3.8: P-values for Hardy-Weinberg equilibrium tests using the two-population baseline datasets with non-linear dependence between f and the covariate. All 10 replicates are shown.

	My method p-value		Genepop p-values	
	for X_{ij}	for D_{ij}	for Heterz. Def.	for Prob. Test
Rep 1	0.038	0.000	<.0001	<.0001
Rep 2	0.036	0.000	<.0001	<.0001
Rep 3	0.002	0.000	<.0001	<.0001
Rep 4	0.260	0.000	<.0001	<.0001
Rep 5	0.010	0.000	<.0001	<.0001
Rep 6	0.082	0.000	<.0001	<.0001
Rep 7	0.002	0.000	<.0001	<.0001
Rep 8	0.012	0.000	<.0001	<.0001
Rep 9	0.000	0.000	<.0001	<.0001
Rep 10	0.000	0.000	<.0001	<.0001

as reference. From the results in the tables, it can be seen that misclassification rates influence the power to detect allele frequency differences. In the two baseline scenarios for two-population simulations in Table 3.10 and Table 3.11, when 50% of the same whale allele pairs are misclassified, Genepop understandably can not detect allele frequency differences in any of the 10 runs, while my method successfully detects two sources of genetic variation. When a smaller portion of animals are placed in the wrong strata, the Genepop test for allele frequency differences remains quite effective at detecting allele frequency differences.

3.6.3 The Structure program

Another method for detecting population structure is that used by the Structure program [13, 29]. Structure uses Bayesian clusturing methods to assign individuals (probabilistically) to the pre-assigned number of populations (K). The main idea is that individuals of unknown origin can be assigned to populations with respect to genotypic likelihoods [16, 5]. Using allele frequency estimates, these likelihoods of a given genotype can be derived for each population, i.e., estimated ancestry of the sampled individuals. Thus, Structure assigns individuals to a population based on their genotypes under the assumption that each population is under Hardy-Weinberg equilibrium and attempts to find the population groupings that are as close to equilibrium as possible.

To test Structure here, I again use the 150 same-whale random allele pairs of the simulated baseline scenarios, each replicated 10 times and consider $K = 2$ clusters. The MCMC approach is implemented using a chain length of $500,000$ and $50,000$ burn in. Structure

Table 3.9: P-values for Hardy-Weinberg equilibrium tests using the one population simulation with linear dependence between f and the covariate, replicated 10 times, with three rates of strata misclassification.

	My method p-value		Genepop p-values with misclassification		
	for X_{ij}	for D_{ij}	rate=0	rate=1/3	rate=1/2
Rep 1	<.0001	0.532	0.6289	0.5408	0.1335
Rep 2	<.0001	0.400	0.4864	0.9676	0.7378
Rep 3	<.0001	0.858	0.3518	0.9464	0.4288
Rep 4	<.0001	0.408	0.1427	0.2684	0.3958
Rep 5	<.0001	0.004	0.7638	0.2363	0.1124
Rep 6	<.0001	0.294	0.1492	0.7275	0.7194
Rep 7	<.0001	0.984	0.8534	0.5274	0.9388
Rep 8	<.0001	0.700	0.9906	0.8938	0.7955
Rep 9	<.0001	0.378	0.0510	0.3492	0.1540
Rep 10	<.0001	0.310	0.8581	0.7703	0.8594

Table 3.10: P-values for Hardy-Weinberg equilibrium tests using the two population simulation with linear dependence between f and the covariate, replicated 10 times with misclassification rates of 2/3 and 1/2.

	My method p-value		Genepop p-values with misclassification		
	for X_{ij}	for D_{ij}	rate=0	rate=1/3	rate=1/2
Rep 1	0.018	<.0001	<.0001	<.0001	0.8318
Rep 2	<.0001	<.0001	<.0001	<.0001	0.7140
Rep 3	0.006	<.0001	<.0001	<.0001	0.1390
Rep 4	<.0001	<.0001	<.0001	<.0001	0.4316
Rep 5	0.024	<.0001	<.0001	<.0001	0.5028
Rep 6	0.004	<.0001	<.0001	<.0001	0.6069
Rep 7	<.0001	<.0001	<.0001	<.0001	0.9703
Rep 8	0.054	<.0001	<.0001	<.0001	0.6180
Rep 9	0.012	<.0001	<.0001	<.0001	0.4893
Rep 10	<.0001	<.0001	<.0001	<.0001	0.6526

Table 3.11: P-values for Hardy-Weinberg equilibrium test using the two population simulation with non-linear dependence between f and the covariate, replicated 10 times, with misclassification rates of 2/3 and 1/2.

	My method p-value		Genepop p-values with misclassification		
	for X_{ij}	for D_{ij}	rate=0	rate=1/3	rate=1/2
Rep 1	0.038	0.000	<.0001	<.0001	0.6503
Rep 2	0.036	0.000	<.0001	<.0001	0.0332
Rep 3	0.002	0.000	<.0001	<.0001	0.1341
Rep 4	0.260	0.000	<.0001	<.0001	0.8562
Rep 5	0.010	0.000	<.0001	0.0004	0.2975
Rep 6	0.082	0.000	<.0001	<.0001	0.5791
Rep 7	0.002	0.000	<.0001	<.0001	0.8207
Rep 8	0.012	0.000	<.0001	<.0001	0.5131
Rep 9	0.000	0.000	<.0001	<.0001	0.1854
Rep 10	0.000	0.000	<.0001	<.0001	0.4684

treats each individual as a cluster and combines the clusters that have the greatest shared ancestry. Thus, unlike the method tried above, Structure assigns an ancestry estimate, or an estimated cluster membership for each individual animal. I will use the percentage of correct classification of individuals to clusters as a measurement of success.

Table 3.12-Table 3.14 display the average misclassification rates for the 10 replicates of the three baseline scenarios. For the two population scenario, the median correct classification rate for the 10 replicates was 89%. At first this sounds like a very good result but, with respect to my goal of distinguishing the sources of genetic correlation, Structure can not distinguish the effects of stock structure and the covariate.

For a visual illustration, a graph of individual ancestries for one of the 10 replicates of each of the two-population baseline scenarios is given in Figure 3.14 and Figure 3.15. These graphs are also known as Distruct graphs. Distruct is a graphical display of subpopulation assignment probabilities for the sampled individuals using the output from a Structure analysis [41]. As can be seen from Figure 3.14 and Figure 3.15, Distruct uses a color scheme to distinguish the subpopulations and uses bars to display the membership coefficient of individuals. In both of the figures, a majority of the bars are predominantly one color or the other; only a few bars indicate unclear ancestry. These plots can be evaluted with the assistence of the output from Structure which gives membership coefficients of each individual for the assigned cluster and the inferred cluster. For Figure 3.14, only 4% of the bars indicate assignment to a subpopulation different from the correct one. For Figure 3.15, this reduces to 3.3%. Thus, for both graphs Structure signals more than one subpopulation resulting in population structure. Although [51] suggests that Structure has

Table 3.12: P-values for Hardy-Weinberg equilibrium tests using my method and misclassification rates using Structure for the one-population simulation with linear dependence between f and the covariate, for 10 replicate runs.

	My method p-value		Structure Misclassification	
	for X_{ij}	for D_{ij}	pop 1	pop 2
Rep 1	<.0001	0.532	0.500	0.500
Rep 2	<.0001	0.400	0.500	0.500
Rep 3	<.0001	0.858	0.500	0.500
Rep 4	<.0001	0.408	0.500	0.500
Rep 5	<.0001	0.004	0.500	0.500
Rep 6	<.0001	0.294	0.500	0.500
Rep 7	<.0001	0.984	0.500	0.499
Rep 8	<.0001	0.700	0.500	0.499
Rep 9	<.0001	0.378	0.500	0.500
Rep 10	<.0001	0.310	0.500	0.499

Table 3.13: P-values for Hardy-Weinberg equilibrium tests using my method and misclassification rates using Structure for the two-population simulation with linear dependence between f and the covariate, for 10 replicate runs.

	My method p-value		Structure Misclassification	
	for X_{ij}	for D_{ij}	pop 1	pop 2
Rep 1	0.018	<.0001	0.048	0.085
Rep 2	<.0001	<.0001	0.069	0.053
Rep 3	0.006	<.0001	0.031	0.040
Rep 4	<.0001	<.0001	0.043	0.027
Rep 5	0.024	<.0001	0.076	0.064
Rep 6	0.004	<.0001	0.054	0.079
Rep 7	<.0001	<.0001	0.102	0.068
Rep 8	0.054	<.0001	0.064	0.047
Rep 9	0.012	<.0001	0.063	0.047
Rep 10	<.0001	<.0001	0.068	0.062

Table 3.14: P-values for Hardy-Weinberg equilibrium tests using my method and misclassification rates using Structure for the two-population simulation with non-linear dependence between f and the covariate, for 10 replicate runs.

	My method p-value		Structure Misclassification	
	for X_{ij}	for D_{ij}	pop 1	pop 2
Rep 1	0.038	0.000	0.070	0.106
Rep 2	0.036	0.000	0.077	0.091
Rep 3	0.002	0.000	0.062	0.059
Rep 4	0.260	0.000	0.076	0.095
Rep 5	0.010	0.000	0.134	0.120
Rep 6	0.082	0.000	0.060	0.090
Rep 7	0.002	0.000	0.061	0.067
Rep 8	0.012	0.000	0.040	0.045
Rep 9	0.000	0.000	0.050	0.042
Rep 10	0.000	0.000	0.052	0.037

weak power and can give misleading results, I did no observe that with my data.

Figure 3.14: Results from Structure (for K=2) for seventh replication of two-population simulations with linear dependence.

It should be noted that GENEPOP, Structure and my method each have strengths and weaknesses with respect to what is trying to be achieved. Although my method does not provide classification of individuals to populations of membership, it has the advantage of being able to distinguish genetic correlation due to a covariate effect and that due to a Wahlund effect.

Figure 3.15: Results from Structure for (for K=2) fifth replication of two-population simulations with non-linear dependence.

Chapter 4

Applications to Real Data

In Chapter 3, I studied how my method performed on simulated datasets and compared it with other related methods. In this chapter I show some applications to real data. I consider two different datasets. The first analysis relates to possible population structure in the Bering-Chukchi-Beaufort stock of bowhead whales, which is of interest to the International Whaling Commission. I use my method to determine if there exists any substock structure in this population and if certain covariates are associated with such population structure. The second analysis pertains to a dataset about black-tailed prairie dogs in the Central Plains Experimental Range and the Pawnee National Grasslands in Weld County, Colorado. I investigate the population structure of these prairie dogs living in colonies and compare results with other research on the same dataset.

4.1 Bowhead Dataset

Bowhead whales are an important natural resource used by Native Alaskans for subsistence and as part of their cultural heritage. The bowhead whales studied here live in Bering, Chukchi, and Beaufort Seas regions of the north Pacific and arctic. Figure 4.1 is a map of this region and the predominant migratory route (see Richardson [60] and Moore *et.al* [52]). As a result of commercial whaling from 1848-1914, bowhead whales became a highly endangered species, but in recent years their numbers are steadily rising [17, 27]. Recently, the International Whaling Commission has identified the importance of detecting the population structure of bowhead whales in this region in order to better formulate hunting regulations.

As illustrated in Figure 4.1, bowhead whales are believed to migrate to the north to Beaufort Sea in the spring and migrate back to the south in the fall [23]. Various scenarios have been suggested about their spatio-temporal migration pattern as well as their genetic population structure. In order to investigate population structure, a genetic microsatellite dataset has been developed [14]. These genetic data are obtained almost exclusively from samples from whales killed during the annual fall and spring aboriginal subsistence harvests. Because this hunting occurs only at a handful of remote villages and only at certain times of the years, the spatio-temporal distribution of samples is very poor and uneven. In the most recent dataset, there are 173 samples of which 148 are from the Barrow region. Table 4.1 displays a summary of the data with respect to the locations where whales were captured (Barrow, Gambell or Savoonga). Figure 4.1 maps the area where these whales

Table 4.1: Summary of the location where bowhead whales were captured within 1995-2005, for samples in my dataset.

Barrow	Gambell	Savoonga
148	9	16

are sampled. Most of the bowhead whales were captured in Barrow region.

My analysis is focused on detecting substock structure when bowhead whales migrate south, past Barrow region during the fall migration season and north, past Barrow region during the spring migration season. I analyze the two datasets that consist of 112 fall Barrow bowhead whales and 36 spring Barrow bowhead whales, each with 22 loci.

Although the conventional wisdom has been that Bering-Chukchi-Beaufort bowheads constitute a single genetically well-mixed stock, a variety of alternative hypotheses have been suggested. The Bering-Chukchi-Beaufort bowheads winter in Bering Sea and pass whaling villages such as Barrow on their way north. It is unclear how many whales spend time in the Chukotka region in spring. The whales summer in the Beaufort Sea until fall when they head southward again. If it is assumed that bowhead whales intermix on comman winter breeding grounds but follow several diverse migratory routes during the rest of the year, then the single-stock hypothesis holds. But different migratory routes and spatial segregations may also represent geographically distinct substocks, in which case the alternative hypothesis of more than one stock could hold [14].

If certain multi-stock hypotheses were true, the time of capture time during the fall at

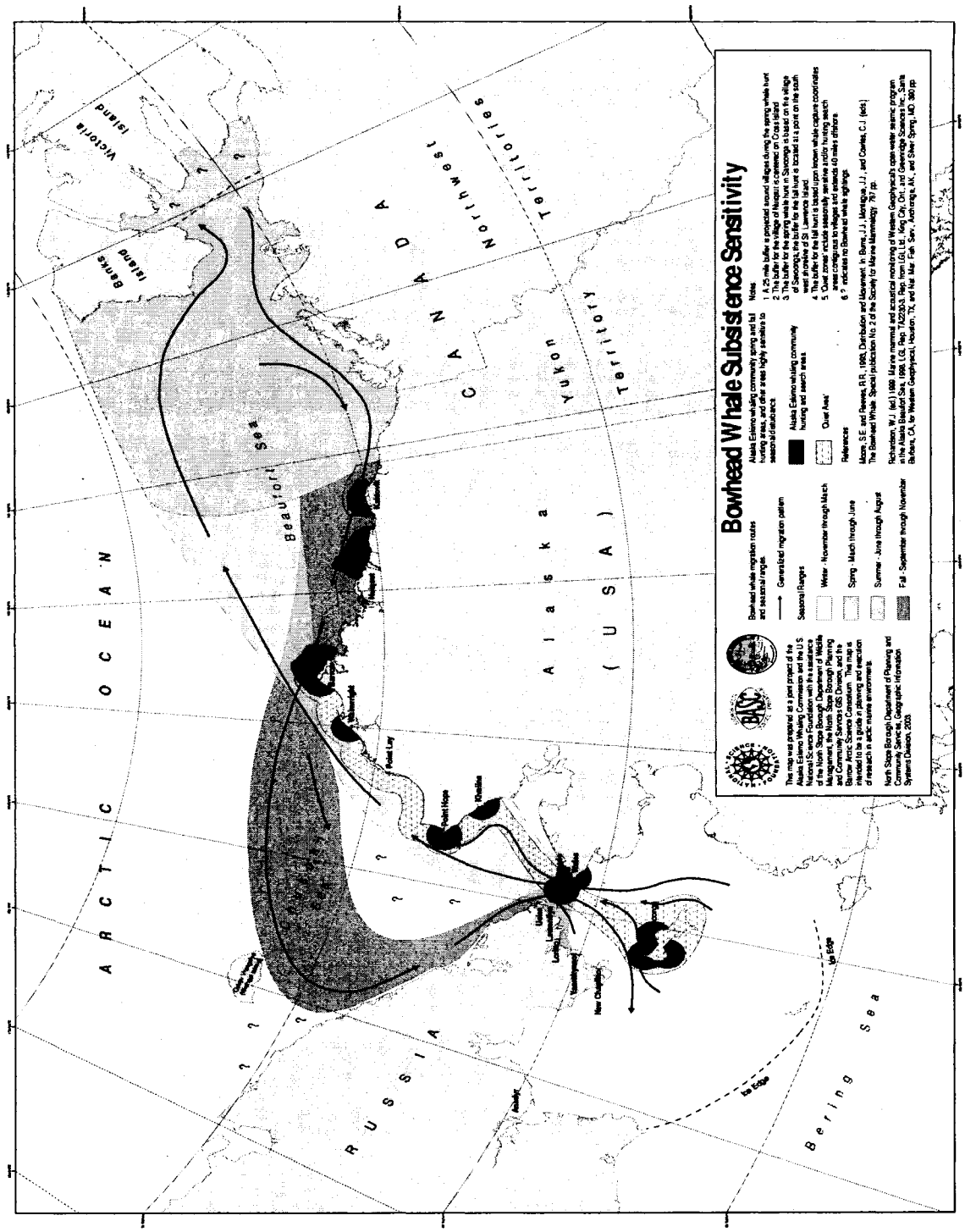

Figure 4.1: Map of possible migration route for Bowhead whales in the Bering-Chukchi-Beaufort Seas during different seasons (www.north-slope.org/nbs/acmp/)

Table 4.2: Numbers of bowhead whales captured at the Barrow region during the fall season. The first and last capture day within each fall season at Barrow is given in the final columns.

	Number of captures	First capture day	Last capture day
1995	4	248	290
1996	18	254	270
1997	20	228	294
1998	1	267	267
1999	6	282	286
2000	7	270	277
2001	4	281	282
2002	19	273	298
2003	5	281	287
2004	15	262	297
2005	13	274	275

Barrow could have strong relationship with allele frequencies. Thus, capture time is potentially an important variable in my analysis. Table 4.2 shows the yearly number of bowhead whales captured at Barrow as well as the first and last capture day within the fall season of that year. The range of capture time of the samples is no more than 66 days during a year and no more than 70 days overall. The difference in first capture time from one year to the other is mainly due to change of first day of fall hunting season and to year-to-year variation in migration timing. The fall season capture times are generally centered around the month of September.

Table 4.3 shows the yearly number of bowhead whales captured at Barrow as well as

Table 4.3: Numbers of bowhead whales captured at the Barrow region during the spring season. The first and last capture day within each fall season at Barrow is given in the final columns.

	Number of captures	First capture day	Last capture day
1995	4	126	152
1996	2	145	150
1997	8	124	155
1998	1	143	143
1999	8	118	143
2000	5	115	151
2001	8	118	134

the first and last capture day within the spring season of that year. The range of capture time of the samples is no more than 36 days during a year and no more than 40 days overall.

Data were collected from tissue samples of the bowhead whales that were captured. Genetic data consist of 11 and 22 variable microsatellite loci [14, 22]. There are two distinct datasets because results found with the first dataset of 11 loci (chosen opportunistically) motivated a 2-year research effort to develop an improved dataset of 22 new, arguably superior loci. While [1] found significant heterozygote deficiency in several of the loci for the first dataset, I was not able to find such an effect for the second dataset. The lengths and capture times of the whales are also provided in both datasets.

4.1.1 Corroboration of Results from Jorde *et al.* [42] using 11 loci dataset

Using 11 different loci on Bering-Chukchi-Beaufort Seas bowhead dataset of 85 whales, Jorde *et al.* [42] found genetic similarity to be less between whales captured in the same year 5 to 11 days apart than between those captured less than 5 or more than 11 days apart. Although a biological explanation for this finding seems elusive, it is possible that such a pattern could be induced by the temporally staggered migration of two genetically distinct stocks past Barrow. I will use difference in capture time as my covariate when applying my method to detect population structure using the bowhead whale datasets described above. The exact dates of migration vary somewhat from year to year due to various factors such as weather, ice conditions, and prey availability. Therefore I will base my analysis on all possible pairwise comparisons of whales within the season and will not focus on a certain group of whales with respect to fixed dates nor make cross-season comparisons of genetic similarity.

Table 4.4 displays the p-values when my method is applied to the 11 loci fall and spring bowhead dataset. Results for the fall 11 loci dataset are consistent with the findings of Jorde *et al.* [42] in that there exists strong evidence of both significant covariate effect and substock structuring. Figure 4.2 and 4.3 show the results in Table 4.4. In each plot the intercept of the two solid lines are different enough to signal substock structuring. The findings by Jorde *et al.* [42] of significant temporal effect is supported in Figure 4.2 with the solid curve going outside of the dotted 95% joint confidence band. The insignificant

Table 4.4: P-values from my analysis of 11 loci fall and spring Barrow bowhead whales.

	p-value		
	for X_{ij} using deviance	for X_{ij} using bands	for D_{ij} using deviance
Fall dataset	<0.001	0.034	<0.001
Spring dataset	0.176	0.086	0.010

covariate effect for the spring dataset is seen in Table 4.3 by the solid curve lying within the 95% joint confidence band.

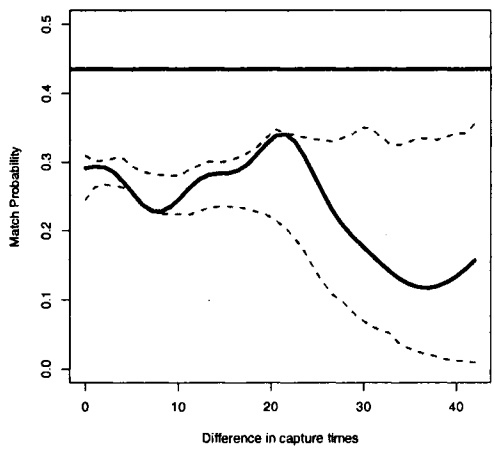

Figure 4.2: Fitted P[match] for the 11 loci fall bowhead whale examples, with covariate being the difference in capture times.

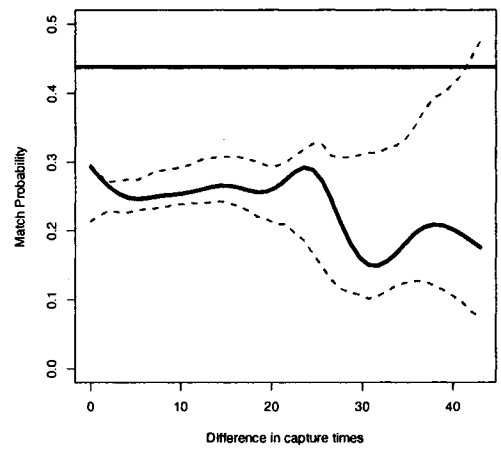

Figure 4.3: Fitted P[match] for the 11 loci spring bowhead whale examples, with covariate being the difference in capture times.

4.1.2 Application using 22 loci dataset

The Bering-Chukchi-Beaufort Seas bowhead dataset of interest consists of 22 different loci on 112 fall and 36 spring bowhead whales. My analysis used 500 permutations, employing the testing methods described in Chapter 2. The resulting p-values are given in Table 4.5. For both datasets, I found non-significant p-values indicating no evidence for substock structure, and non-significant p-values for the covariate effect indicating no significant temporal pattern. Figures 4.4 and 4.5 show my findings by plotting the fitted match probability using the model given in equation (2.49) for the two datasets. In both plots, the intercepts of the two solid lines are similar enough to suggest no significant substock structure, while the different whale pairing curve is within the 95% confidence band indicating no significant covariate effect and thus no significant temporal pattern.

Thus, my findings are entirely negative. This is an important contrast to the findings of [42]. A possible explanation for this difference is that [42] used 11 loci on 85 whales, whereas I used 22 different (and arguably superior) loci on 112 whales (including many of the same 85). Indeed, when my analysis is repeated using only the whales and loci analyzed by Jorde *et al.*, as shown in Section 4.1.1, significant results are found. Another possible reason for such a contrast to the findings of [42] is allele dropout occuring for loci and primers not designed specifically for bowheads. As a result of allele dropout, the number of homozygotes increases, signaling genetic correlation structure. Thus, if allele dropout has occured in the 11 loci dataset, findings by [42] may have been caused by allele dropout and not genetic correlation structure. For the 22 new loci developed specifically for bowheads,

Table 4.5: P-values from my analysis of 22 loci fall and spring Barrow bowhead whales.

	p-value		
	for X_{ij} using deviance	for X_{ij} using bands	for D_{ij} using deviance
Fall dataset	0.084	1	0.376
Spring dataset	0.728	1	0.274

the likelihood (and apparent frequency) of allele dropout is much lower, and no genetic correlation structure was found in my analysis.

Overall, my findings indicate that there is no substock structuring. That is, there appears to be a single stock of bowhead whales migrating in the region of the Bering-Chukchi Beaufort Seas. There is no within-substock genetic correlation structure related to patterns of temporal separation. Under these circumstances, it is of interest to check whether related tools such as GENEPOP and Structure are able to detect any substructure.

The GENEPOP results for the 22 loci fall and spring bowhead datasets are given in Table 4.6. GENEPOP results are consistent with my results of no source of Hardy-Weinberg disequilibrium.

When each bowhead whale was preassigned to a single population, results from another related method, Structure [13, 29], are given in Table 4.7. As mentioned in previous

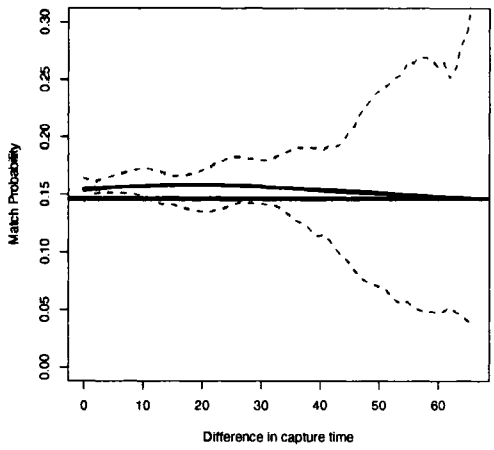

Figure 4.4: Fitted P[match] for the 22 loci fall bowhead whale examples, with covariate being the difference in capture times.

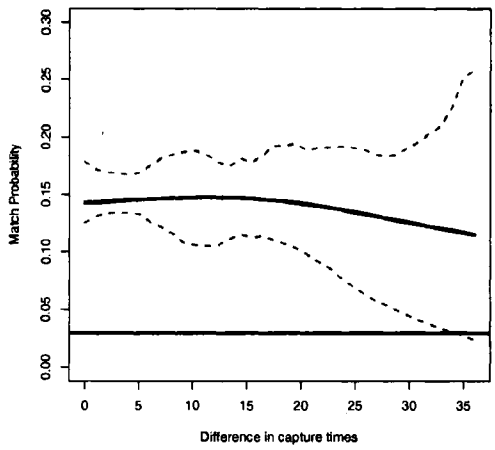

Figure 4.5: Fitted P[match] for the 22 loci spring bowhead whale examples, with covariate being the difference in capture times.

chapters, Structure probabilistically assigns sampled individuals to predetermined subpopulations using a Bayesian clustering approach. The corresponding DISTRUCT [41] plot is

Table 4.6: P-values for Hardy-Weinberg disequilibrium tests using GENEPOP for the fall and spring Bering-Chukchi-Beaufort bowhead whales captured at the Barrow region.

	GENEPOP p-values		
	for Heterz. Def	for Prob Test	for Hetrz. Excess
Fall dataset	0.11	0.47	0.21
Spring dataset	0.38	0.11	0.62

Table 4.7: Inferred population rates using Structure for the 22 loci fall and spring Bering-Chukchi-Beaufort bowhead whales captured at the Barrow region.

Structure Inferred Population Rates		
	pop 1	pop 2
Fall dataset	0.502	0.498
Spring dataset	0.501	0.499

given in Figures 4.6 and 4.7. The results in Table 4.7 show inferred population rates. These rates provide no evidence against the single-stock hypothesis, supporting the results from fitting my model to the 22 loci fall and spring Bering-Chukchi-Beaufort bowhead whale stock captured at the Barrow area. I conclude that my method results are consistent with those of GENEPOP and Structure.

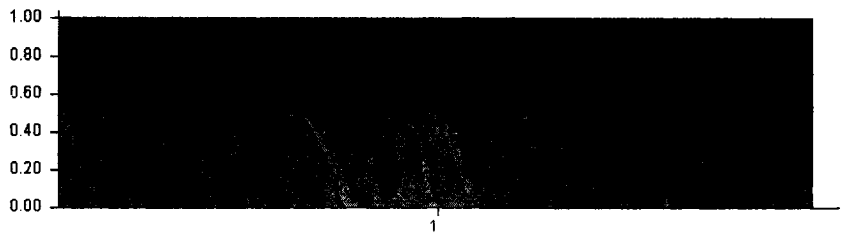

Figure 4.6: DISTRUCT plot for 22 loci fall bowhead whale dataset.

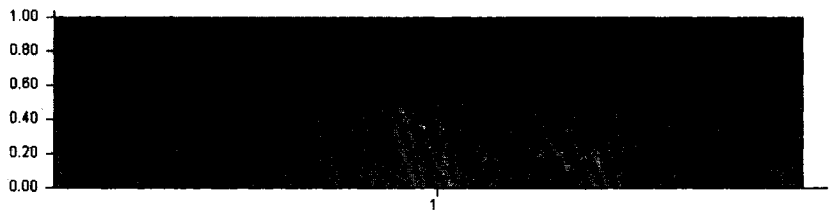

Figure 4.7: DISTRUCT plot for 22 loci spring bowhead whale dataset.

4.2 Prairie Dog Dataset

The prairie dog dataset consists of 155 individuals from 13 colonies in the regions of the Central Plains Experimental Range and the Pawnee National Grasslands in Weld County, Colorado. Data were provided by [30]. A summary of the dataset is given in Table 4.8. Among the 155 individuals, 74 are from colonies in the Pawnee National Grasslands (with colony labels 66-81) while the remaining 81 individuals are from Central Plains Experimental Range (with colony labels 5-29). Along with the number of individuals in each colony, the number of years since recolonization and the area of each colony are given in Table 4.8. There is an obvious correlation between colony age and colony size: colony size tends to increase with increasing colony age.

Genetic data for 7 loci are provided for each sampled individual from each colony. Data were collected from the $264km^2$ area by trapping 3-16 prairie dogs per colony and taking tissue samples. Considering the influence of landscape structure, various measurements were taken to describe the physical distance between colonies. Among these, distance between colonies along drainages was found to be the most important in a previous analysis of these data [30]. Figure 4.8 maps the region where colonies are located. Roads and drainages which are used to compute physical distance measures are also shown on the map.

Previous research had reported moderate levels of substock structuring between colonies. [30] modeled various genetic distance measures, one of which was $F_{ST}/(1 - F_{ST})$, using a covariate as an independent variable. When both age of colony and drainage distance be-

Table 4.8: Summary of the prairie-dog dataset.

Colony	Number of individuals	Colony age (years)	Colony size (ha)
PNG	3	1	1.0
CPER	16	1	3.1
CPER	11	1	3.8
CPER	10	2	2.2
CPER	15	2	2.4
CPER	14	2	2.8
CPER	15	2	6.1
PNG	15	4	4.0
PNG	12	4	7.6
PNG	10	4	7.9
PNG	8	4	18.0
PNG	11	8	31.9
PNG	15	10	52.0

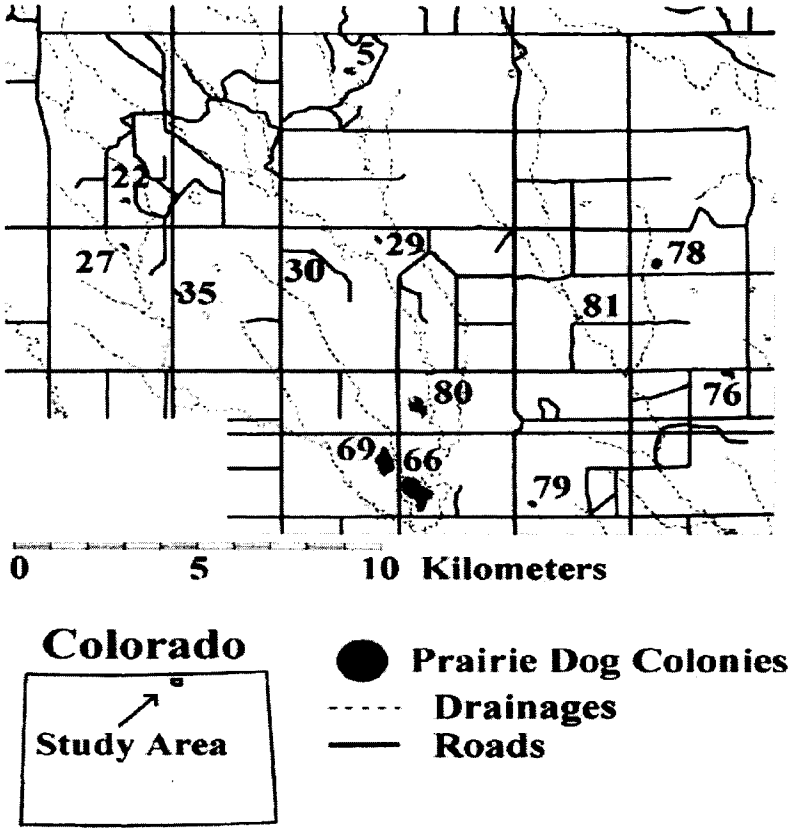

Figure 4.8: Map of 13 black-tailed prairie dog colonies, roads and drainages in the north central part of Colorado. The colonies with site labels less than 50 are in the Pawnee National Grasslands region while those with labels greater than 50 are in the Central Plains Experimental Range region. The range of each colony is shaded on the map next to the colony site label.

tween colonies were used to predict genetic distance, significant relationships were found [30].

As a result, [30] concluded that dispersal occurs among prairie dog colonies after re-colonization. Moreover, this explained the relationship between colony age and genetic distance. That is, earlier recolonized colonies are more genetically distant from each other but with time, they become more similar.

My analyses of these data are designed to investigate the same questions addressed by [30], using a different methodology. If my methods find results that are corroborate earlier findings, this would serve to build confidence in my approach and would support the work of those earlier researchers. Unfortunately, the covariates in the dataset for my analysis relate to colonies, not to individuals (prairie dogs). This means that pair of individuals in the same colony will be assigned the same covariate value. This is different than the simulations in Chapter 3 and my analysis of the bowhead dataset.

I consider two versions of my model, each using a different predictor. In the first analysis, I use age differences (among colonies, as determined by recolonization time) as my covariate to explain the population structure of the prairie dog population. My second analysis employs mean age of colony pairs as the covariate.

If recolonization time has an influence in how much colonies are genetically distant, then by using the difference in colony age as my covariate, my model will find significant covariate effect. [30] used age as one of the variables predicting genetic distance. I will

show that my method is flexible enough to use not only covariates in the form of difference measures but also as any other function; in this particular case as the mean function.

As mentioned before in Section 3.1, when the covariate is a measure of pairwise differences, it takes the value zero for all same individual allele pairs. For the black-tailed prairie dog application, this is still valid when the covariate is age difference among colonies but does not hold for the case of mean age. Since colony ages vary, the mean age of the same individual allele pairs will have the age of the colony as the covariate value. In match probability plots for the bowhead dataset, the same-whale match probability was a single point at $X_{ij} = 0$ with dotted lines shown for reference. This is still the case when the covariate is age difference among colonies for the black-tailed prairie dogs but consists of more than one point when the covariate is mean age. These points consist of same individual allele pairs match probabilities at the covariate value of the colony ages.

This change in same-individual match probability also requires changes to the model fitting and hypothesis testing approaches described in Section 2.3.3-2.3.4. After some adjustments to equation (2.51) (considering the varying same-individual paired allele match probabilities), I fit the model in equation (4.1) for the case when the covariate is the mean age of the colony.

$$Z_{ij}^{\ell} = \mu + \gamma_0 \delta_{ij} + s(X_{ij}) + \beta_{\ell} + \gamma_{\ell} \delta_{ij} + s_{\ell}(X_{ij}) \tag{4.1}$$

respectively, where $\sum \beta_{\ell} = \sum \gamma_{\ell} = \sum s_{\ell} = 0$ and X_{ij} is the mean age of the corresponding colonies of the ij^{th} allele pair. When allele pairs are from the same-colony, $\delta_{ij} = 0$ and

therefore equation (4.1) reduces to

$$Z_{ij}^{\ell} = \mu + s(X_{ij}) + \beta_{\ell} + s_{\ell}(X_{ij}). \tag{4.2}$$

If there exists a single stock, then the probability of two alleles from different colonies should equal that for alleles from the same colony, i.e., $\gamma_0 = 0$. Thus, testing the single stock hypothesis is equivalent to testing $H_o : \gamma_0 = 0$.

Testing for the covariate effect is equivalent to testing the significance of s because $s(X)$ should be flat if the covariate has no affect on match probability. Thus, testing for the significance of the covariate is carried out using deviance tests and null confidence bands as explained in Section 2.3.4.

The resulting p-values after fitting my model using (2.51) when age difference is the covariate and using (4.1) when mean age is the covariate, are given in Table 4.9. There is significant covariate effect when either difference in recolonization time or mean age of paired colony is used as a covariate. Both models indicated significant covariate effect as well as when the confidence band method was used to find the p-value. Moreover, for both cases, my model indicates no significant substock structure aside from the structure associated with the covariates.

The fitted match probabilities for the two analysis are graphed in Figure 4.9 and Figure 4.10. The upper solid line is the case for individual pairs from the same colony, where this line is for reference in Figure 4.9 representing a single point. The curved lower solid line is the fit for different individual allele pairings, from different colonies with 95% confidence

Table 4.9: P-values for tests of population structure and covariate effects using my method for the black-tailed prairie dogs in the Pawnee National Grasslands and Central Plains Experimental Range regions of Weld County, Colorado.

	My Method p-value		
	for X_{ij}	for X_{ij}	for D_{ij}
covariate	using deviance	using bands	using deviance
Difference in recolonization	0	0	0.946
Mean age of the colony pair	0	0.002	0.948

bands shown with dotted lines.

In both plots in Figure 4.9 and Figure 4.10, the curve lines go outside of the 95% confidence bands, indicating significant covariate effects. Thus both difference in recolonization time and mean age of colony pair have a significant association with genetic structure. Since same-individual match probability is a single point when the covariate is difference in recolonization time, the gap between the two solid lines at the intercept of Figure 4.9 is a measure of substock structuring. When the covariate is mean age, same-individual match probability is no longer a single point at $X_{ij} = 0$ but consist of a curve parallel to the different-individual match probability curve. Therefore, the difference between these two curves is a measure of substock structuring. Table 4.9 indicates that Figures 4.9 and 4.10 do not constitute evidence of genetically distinct substocks.

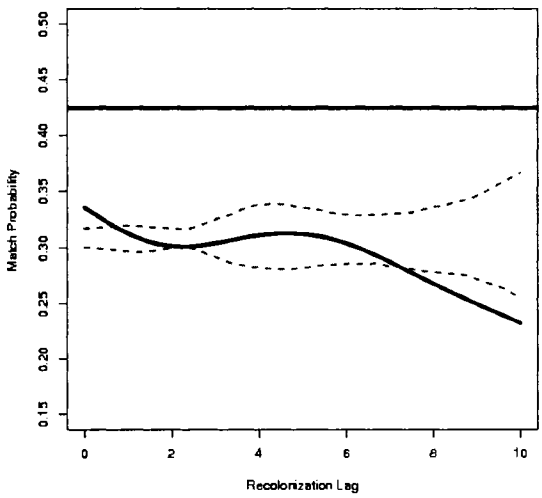

Figure 4.9: Fitted model when covariate is difference in recolonization time

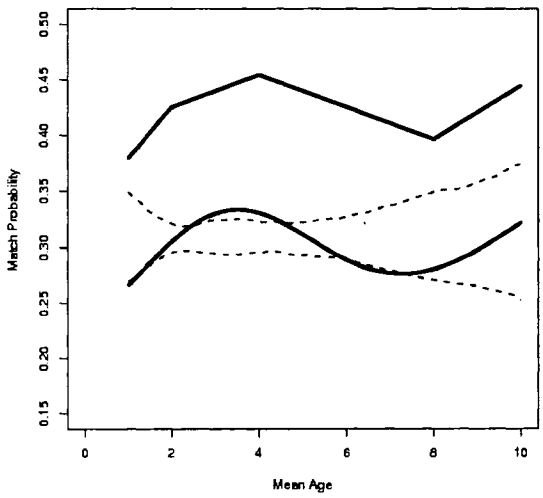

Figure 4.10: Fitted model when covariate is mean age of paired colony

As a result of applying my method to the black-tailed prairie dog dataset, I have found significant covariate effect in both cases where I assumed within substock genetic structure is associated with difference in recolonization time or mean age. This is consist with the findings of [30] who found significant association between genetic distance and ages of the colonies. That is, older colonies are more genetically similar than younger colonies. Moreover, [30] stated that over time, colonies become more genetically similar due to dispersal among colonies.

In order to test population structure of black-tailed prairie dogs, I have used hypothesis testing and found no significant population structure. This result is consistent with the findings of [30] in that no severe population structure was found.

Chapter 5

Summary, Future Work and Conclusions

In the previous chapters, I have developed a new method that models and tests the genetic correlation structure of a population using microsatellite data. My model relates within substock correlation structure to a covariate and detects genetic differentiation of distinct substocks. My simulation results and applications to real data suggest that the method can be useful in a variety of situations and has sufficient power to detect scientifically interesting effects. Here I will make some general conclusions about this model and make recommendations about what further methodological enhancements might be useful.

5.1 Summary

In this dissertation I have developed a new method to model the genetic structure of a population. A variety of methods have been developed to detect within or between substock genetic structure, and I view my approach as an addition to this collection of tools, not as a replacement for other methods. In my method I distinguish effects between and within substocks and develop valid tests that determine their significance. My method is flexible enough to be applied in various areas of research such as population genetics, forensic science, and ecology, and is compatible with different types of genetic data as well as multiple covariates. As with the black-tailed prairie dog example, the influence of environmental, biological, and ecological factors can be tested using different covariates.

The method developed in this dissertation has made several contributions to statistical genetics and population genetic science. These contributions can be summarized as:

1. The modeling approach was based on the paired alleles in the population. Difference between same-individual allele pairings and different-individual allele pairings allow estimation of population substructure. Although a few paired-individual approaches have been suggested previously [39, 40, 42], none have the same degree of individual-based allele matching. My approach represents a new strategy for modeling.

2. Unlike some other related methods, my method does not require user-specified speculative stratification, nor does it require the number of substocks. While my approach does not provide an estimated stratification, it does use the data to detect the effects

of strata. In some cases, my method might be adjusted to test for the number of substocks after detecting multiple substock structure through the use of indicator variables as covariates to identify specific groupings. In this dissertation I have considered the overall Wahlund effect and have not considered estimating the number of substocks influencing the occurence of Wahlund effect.

3. The within-substock correlation structure was related to a measurable covariate in my model. Each allele pair had a corresponding covariate. Relating genetic variation to covariates is usually done at the population level (e.g., [30]) and only rarely has been attempted at the individual level (e.g.,[54]). The derivation of developing a new testing procedure to test within substock correlation structure using this covariate was given in Chapter 2, Section 2.3.4.

4. Analysis of real datasets revealed important biological conclusions about population structure. In the application of black-tailed prairie dogs, we confirmed earlier findings suggesting that recolonization time is associated with genetic similarity. In the case of bowhead whales, when my method was applied to the 11 loci fall bowhead dataset we found results confirming the results of Jorde *et al.* [42]. The analysis results for the 22 loci fall and spring bowhead datasets show no evidence of population structure or covariate effect, which is an important conclusion for wise management.

5.2 Future Work

There are many factors that influence the genetic structure within a population. When developing this new method, I made some basic assumptions including: no genetic migration; mutation is not occurring; no natural selection; equally likely and well-mixed breeding, etc. In real life some of these assumption may not hold.

In this section I will go over two important areas where I made simplifying assumptions that merit further consideration: the effect of sibship and the influence of linkage. I also consider application using different genetic data, particularly Single Nucleotide Polymorphisms (SNP's), and the case with mulitiple covariates. These are areas for further research, and possible adjustment to various aspects of modeling and testing that was used in this dissertation. After giving definitions of what these factors are and their influence on the dynamics of a population, I give recommendations on how to approach these cases to include their impact on my method.

5.2.1 The Sibship Effect

When two individuals share a mother or father they are said to be siblings. Having siblings in a population has an effect of increasing the proportion of individuals that share similar genes. Therefore, when alleles are randomly chosen from two individuals, the probability of getting a match is higher when the individuals are siblings than when they are

not. Thus, it is important to consider the effect of siblings when modeling match probabilities. In the bowhead whale application, the population is large (estimated 10,585 in 2001 [27]) relative to the sample size of 112, however sibships within the dataset cannot be wholly discounted because (i) some whales may travel in loose familial groups, and (ii) the population has recovered from a very small size since 1914 yet estimated bowhead lifespans can exceed 150 years [31]. In the prairie dog dataset, sibships are also a possible question. Colony sizes are moderate and spatially isolated. Although black-tailed prairie dogs reproduce at a much faster rate with respect to bowhead whale, they consist of metapopulations which become extinct after plague episodes. Thus, the influence of sibship might have an influence on genetic population structure, but this influence might not be as significant due to the metapopulation property of black-tailed prairie dogs.

In order to measure the performance of my method when there exists siblings, one would naturally turn first to simulation studies. In Chapter 3 I simulated individuals as pairs and thus did not have complete genetic data for each individual. The reason for this approach was to create gene pools that mimic the real-life case under the assumption of no linkage, using a far simpler simulation strategy permitted by that assumption. The dataset consisted of pairs of alleles from the paired individuals, of only a single locus per case. The corresponding covariate was also generated for the pair, not for the individuals themselves. Although these shortcuts simplified and speeded up the computation while still providing the correct genetic frequencies and correlation structure under the no-linkage assumption, they had the disadvantage of not providing complete data for each individual in the dataset.

Below I describe two approaches to creating genotypic data for offsprings. As I will

show, problems arise not in the simulated breeding, but instead in attempting to create allele pairings, and genetic structure related to the covariate.

(1) It is possible to create certain offspring from the current simulated data. When same-individual allele pairs were simulated, the two alleles from each parent are simulated and known. By crossing the two parent genotypes, offspring siblings genotype can be simulated.

Next, the siblings would need to be paired with every other individual in the dataset. Since the covariate is (typically) a measure of the difference between covariate values for the individuals constituting the pair, same-individual allele pairs are assigned covariate values of zero. The first problem now arises: do offspring inherit their covariate values from their parents. If not, this approach is stymied. The next problem arises when the offspring are paired with every other individual. What covariate values should the pairs be assigned? This is crucial since the covariate is supposed to be related to the correlation structure within the substocks. In order to accurately measure the influence of siblings these covariates need to be accurate. In the case of using same-individual pairs, this is not possible.

(2) It is also possible to create offspring siblings from the different-whale simulated data. When different individual pairs from different populations were simulated, each individual was first assigned a genotype. Next, an allele from each genotype was chosen and the binary variable Y was assigned a value depending on whether or not the two alleles matched. Table 5.1 shows an example of how single-locus data for two allele pairs (where the alleles originated from separate substocks in this case) is represented in the dataset. Notice that the genotypes for the two individuals are available.

Table 5.1: Portion of simulated dataset for two allele pairs where in both cases the alleles originated from different populations. The first two columns are for the individual contributing the first allele in the pair and the 3^{rd} and 4^{th} columns are for the individual contributing the second allele in the pair. POP represents the population of origin for each individual and GEN represents the genotype of the corresponding individual. ALLELE 1 is the randomly chosen allele from the genotype in the 2^{nd} column while ALLELE 2 is the randomly chosen allele from the genotype in the 4^{th} column. If ALLELE 1 and ALLELE 2 match, Y is assigned 1, therewise it takes the value zero. The covariate for the allele pair is randomly assigned from a Uniform (0,1) distribution. Of course, the genotypes are simulated after the covariate has been drawn, in order to ensure the desired genetic structure. Only the covariate and Y are needed to fit my model.

POP	GEN	POP	GEN	ALLELE 1	ALLELE 2	COVARIATE	Y
POP 1	AA	POP 2	AB	A	B	0.170	0
POP 2	AB	POP 1	BB	B	B	0.014	1

For the case in Table 5.1, offspring can be generated as follows. The genotypes of the parents are determined by choosing the genotypes in the same populations. That is, offspring in POP 1 can be generated by crossing the genotypes AA and BB, while offspring in POP 2 can be generated by crossing AB and AB. Thus, it is possible to create siblings and their genotype. Again, however, the covariate of the allele pairs related to such offspring cannot be found. As a result, it is not possible to pair up the siblings with the individuals and measure their influence.

As can be seen in both of the cases given above, it is possible to create offspring and siblings, and their genotypes. The fundamental problem arises because I have not used an evolutionary model to simulate populations through time, incorporating a mechanism to introduce the covariate effect. I also have not created datasets that literally compare each individual to each other individual at each locus. If the loci are not linked, then separate comparisons between different individuals at three loci are comparable to comparisons between the same two individuals at those three loci.

My avoidance of a full evolutionary model was a practical choice. It is unclear whether any such model exists for simulation of the sort of data needed to test my analysis approach. I am aware of one group of scientists at the National Marine Mammal Laboratory in La Jolla who are attempting to code such a model. Their work is not yet suitably advanced for use here. My simulation approach does provide the correct allele frequencies and genetic correlation structure for the instant in time when the data are sampled. The simulation of these data requires only brief CPU time, which enables the generation of a large number of datasets for testing estimation performance under different scenarios. Unfortunately, offspring and siblings cannot be generated from these data without losing the connection to the covariate.

My decision not to create datasets by comparing each individual with each other individual was also a practical one. In real applications, including those discussed in this dissertation, the intended use of my model is for such all-possible-pairings data. However, the shortcut I employed allowed me to generate genotypes for individuals in a pair because

I needed to worry only about the covariate value for that pair. Thus, I could determine allele frequencies and match probabilities for only that pair of alleles, separately of what was simulated for other allele pairs. Suppose I had instead attempted to generate all-possible-pairings data. Then, the alleles for the first individual would need to be drawn in such a way that they had the correct match probabilities with the alleles from each other individual, based on every other individual's covariate value. The same requirement would hold for each of the other individuals. To ensure all the right correlation structure would be impossible without a full evolutionary model, and even then it is extremely difficult to imagine how to parameterize such a model to yield current genetic structure of exactly the desired nature and strength.

Although it is difficult to envision how to simulate complete genetic data of the desired sort, it is not as difficult to imagine that such structure can exist in real populations. Genetic variation over space is well-known in many populations [42], so if the covariate is related to spatial separation, then models like mine are one natural approach. In the case of the bowhead whale or other populations that have passed through severe bottlenecks, it is possible that a covariate related to age may capture the signal of genetic drift caused by the bottleneck.

5.2.2 The Influence of Linkage

Unlinked loci are independent. If two loci exhibit linkage, the distance between loci on the chromosome can be used as a measure of their correlation or linkage.

One assumption that is made when generating data is that there is no linkage among loci. That is, that loci are not correlated with each other. Using GENEPOP I found no evidence of linkage in the Fall 2006 bowhead whale dataset. The same results hold for the black-tailed prairie dog dataset. Under this assumption of no linkage I was able to generate each locus independently for my simulation studies.

The assumption of no linkage among loci also brought an advantage in how the generalized additive model was fit. Since the loci were considered to be independent, the response outcomes from different loci were modeled as independent. When linkage is present these outcomes are not independent and adjustments to the analysis method would be needed. One solution might be to merge the information in the correlated loci. That is, if two loci are correlated, instead of coding Y as a binary variable, assign it values as follows:

$$Y_{ij} = \begin{cases} 0, & \text{if } i\text{th and } j\text{th alleles do not match in either loci} \\ 1, & \text{if } i\text{th and } j\text{th alleles match in only one of the loci} \\ 2, & \text{if } i\text{th and } j\text{th alleles match in both of the loci.} \end{cases} \tag{5.1}$$

This change in the Y variable would require changing the structure of the generalized additive model and its link function. McCullough and Melder [44], discuss the analysis of polyamous data in generalized models; analogous ideas would pertain to generalized additive models.

An alternative approach suggested by Robin Waples (pers commn) is to pair multi-locus gametes instead of single-locus alleles. In addition to handling such linkage, such an approach might also provide greater statistical power (but also greater complexity).

5.2.3 Application To Other Types of Genetic Data : SNP's

In this dissertation the main focus of application was microsatellite genetic data. My method is flexible enough to be applied to different types of genetic data. Single Nucleotide Polymorphisms (SNP's) are a type of genetic marker that is used for various areas of research. It is defined as a small change in genetic pattern or variation in a person's DNA sequence [18] as follows. A segment of DNA consists of a sequence of nucleotides among A,T,G, and C. A SNP is a DNA sequence variation occurring when a single nucleotide - A, T, C, or G - in the genome (or other shared sequence) differs between members of a species (or between paired chromosomes in an individual). For example, two sequenced DNA fragments from different individuals, AAGCCTA to AAGCTTA, contain a difference in a single nucleotide. In this case there are two alleles: C and T. This is especially important when such a genetic change effects the coding for protein production since it may influence the biological functioning [18].

When SNP is located on a chromosome, the corresponding nucleotides, A, C, G or T, are coded and used as genetic data for that individual. This is done for all individuals in the population. SNP can also be displayed as diploid data by considering the allele pair on

the corresponding locus. In the latter case, my method is applicable by creating pairings of each individual. This provides the opportunity for comparison of the probabilities of SNP allele (i.e. nucleotide) matchings between and within individuals, and for relating the former probability to a covariate. Conceptually, therefore, SNP data can be analyzed using the same framework presented in this dissertation, after the data have been suitably organized.

5.2.4 Multiple Predictors

For simplicity I have applied my method to cases where a single covariate is of interest. My method is flexible enough to be used with multiple covariates. Of course, simple adjustments to the hypothesis testing methods will be needed.

Let X_{ij} and W_{ij} be two covariates that are related to the genetic correlation structure within substock; i.e. $f = \alpha_0 + \alpha_1 s(X_{ij}) + s(W_{ij})$. Let θ and g be defined as previously, i.e. $g = \theta + f - f\theta$. Note that the change in the assumption about genetic correlation structure with respect to the covariates does not affect the match probabilities. Thus, equations (2.43-2.48) still hold. The generalized additive model defined in equation (2.51) is revised to be

$$Z_{ij}^{\ell} = \mu + \gamma_0\delta_{ij} + s(X_{ij})\delta_{ij} + s(W_{ij})\delta_{ij} + \beta_\ell + \gamma_\ell\delta_{ij} + s_\ell(X_{ij})\delta_{ij} + s_\ell(W_{ij})\delta_{ij} \qquad (5.2)$$

where $\sum\beta_\ell = \sum\gamma_\ell = \sum s_\ell = 0$. Note that, since some of the binary responses are dependent, the parameter estimates from fitting this model via GAM are not Maximum

Likelihood Estimates and thus standard testing procedures can not be used. Therefore permutation testing needs to be used to test the parameters.

Testing for population structure and testing the significance of X_{ij} are the same as described in Section 2.3.4. In addition, the significane of W_{ij} needs to be tested. Similar to the procedure for testing W_{ij}, the columns in the dataset related to W_{ij} are shuffled. The reduced additive predictor is $Z_{ij}^{\ell} = \mu + \gamma_0\delta_{ij} + s(X_{ij})\delta_{ij} + \beta_\ell + \gamma_\ell\delta_{ij} + s_\ell(X_{ij})$. If W_{ij} does not influence f, then the test statistic —namely the deviance change between fitting model (5.2) and the reduced additive predictor— should be insignificant. If this test statistic is unusually different from the null distribution, then the null hypothesis is rejected and we conclude that the covariate has a significant effect in explaining the correlation structure. In addition to deviance tests, joint coverage null confidence bands can be used to test the effect of W_{ij}. If the fits go outside these bands than we conclude that W_{ij} is significant.

In Chapter 3, accuracy of the model fits were evaluted by visually comparing the true match probability plots and the fitted match probability plots. These plots can also be used when there are multiple predictors explaining the genetic correlation structure. In addition to these plots, classical diagnostic methods can be used to check for goodness-of-fit, overdispersion or linearity. Perhaps the most important question related to model diagnostics is goodness-of-fit. It is beyond the scope of this discussion to fully address this question, however, there are various methods based on estimating the dispersion parameter for testing over-dispersion [44]. One should be cautious in applying some of these methods in my simulations since the binary responses are correlated.

5.3 Conclusions

In this dissertation, I have developed a new method that detects, models and tests the genetic correlation structure of a population where some within-subpopulation genetic structure can be explained by a covariate. The modeling approach is based on match probabilities of allele pairs (noting whether the source whales are the same individual or not) and does not require pre-specified population identity strata. The significance of effects attributable to population substructure and to the covariate is determined by permutation testing.

With a suitable covariate choice, my method can be used to determine spatial and/or temporal effects. In Chapter 4, a spatial effect was detected using black-tailed prairie dog dataset. When age difference of paired colonies were used as covariate, there was no significant effect, but when mean age of paired colonies were used as covariate, a significant covariate effect was found. The bowhead analysis focused on genetic patterns associated with temporal separation in a migration path.

The method developed in this dissertation is intended to complement —not replace— other methods for detecting population structure. My method has an advantage of not requiring prespecified strata, and permitting inference about the effect of covariates. However, my method does not provide estimated assignments of individuals to strata, nor does it provide a direct means for estimating the number of subpopulations present in the dataset. Simulation results and real data applications shown here suggest that my method had ad-

equate power to be a useful new tool for the detection and analysis of genetic population structure.

Bibliography

[1] R.L.Honeycutt A.P.Rooney and J.N. Derr. Historical population size change of bowhead whales inferred from DNA sequence polymorphism data. *Evolution*, 55:1678–1685, 2001.

[2] A.S.Dyke, J.Hooper, and J.M.Savelle. A history of sea ice in the canadian artic archipelago based on postglacial remains of bowhead whale (balaena mysticetus). *Artic*, 49:235–255, 1996.

[3] C.F.Kercher & Associates. http://www.greenfacts.org/glossary/def/dna.html.

[4] B.S. Balding and R.A. Nichols. DNA profile match probability calculation : how to allow for population stratification, relatedness, database selection and single bands. *Forensic Science International*, 64:125–140, 1994.

[5] B.Rannala and J.L.Mountain. Detecting immigration by using multilocus genotypes. *Proc. National Academic Society, USA*, 94:9197–9201, 1997.

[6] B.S.Balding. Likelihood-based inference for genetic correlation coefficients. *Theoretical Population Biology*, 63:221–230, 2003.

[7] B.S.Weir. *Genetic Data Analysis II.* Sinauer, 1996.

[8] B.S.Weir and C.C.Cockerham. Variance of actual inbreeding. *Theoretical Population Biology*, 23:85–109, 1983.

[9] B.S.Weir and C.C.Cockerham. Estimating f-statistics for the analysis of population structure. *Evolution*, 38:1358–1370, 1984.

[10] B.S.Weir and W.G.Hill. Estimating f-statistics. *Annual Review of Genetics*, 36:721–750, 2002.

[11] C.C.Cockerham. Variance of gene frequencies. *Evolution*, 23:72–84, 1969.

[12] C.C.Cockerham. Analysis of gene frequencies. *Genetics*, 74:679–700, 1973.

[13] M. Stephens D. Falush and J.K.Pritchard. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164:1567–1587, 2003.

[14] D.D.Hunter. Did bowhead whales (balasena mysticetus) from the Bering-Chukchi-Beaufort Seas undergo a genetic bottleneck? A test using nuclear microsatellite loci. MS Thesis, Texas A&M University, August 2005.

[15] D.E.May. ISB 202: Gene and cell replication, applied environemental science and organizational biology. http://www.msu.edu/course/isb/202/ebertmay/notes/snotes/02-07-06-genes1.html, 2006.

[16] D.Paetkau, W.Calvert, I.Stirling, and C.Strobeck. Microsatellite analysis of population structure in canadian polar bears. *Molecular Ecology*, 4:347–354, 1995.

[17] D.Rugh, D.DeMaster, A.Rooney, J.Breiwick, K.Shelden, and S.Moore. A review of

bowhead whale (balaena mysticetus) stock identity. *Journal of Cetacean Research and Management*, 5:267–279, 2003.

[18] National Center for Biotechnology Information. A Science Primer. http://www.ncbi.nlm.nih.gov/About/primer/index.html, 2006.

[19] New Zealand Genetics. Livestock improvement. http://www.newzealandgenetics.com.

[20] G.H.Givens and I.Ozaksoy. Population structure analysis based on pairwise microsatellite allele matching frequencies in the abscence of source whale population information. Paper IWC SC/57/SD1 presented to the Scientific Committee of the International Whaling Commission, May 2005.

[21] G.H.Givens and I.Ozaksoy. Population structure analysis based on pairwise microsatellite allele matching frequencies. Paper IWC SC/57/SD1 presented to the Scientific Committee of the International Whaling Commission, May 2006.

[22] G.H.Givens and I.Ozaksoy. Transience of temporal lag correlation feature in bowhead microsatellites. Paper IWC SC/57/SD1 presented to the Scientific Committee of the International Whaling Commission, May 2006.

[23] G.H.Givens, J.W.Bickham, C.W.Matson, and I.Ozaksoy. Examination of Bering-Chukchi-Beaufort Seas bowhead whale stock structure hypotheses using microsatellite data. Paper IWC SC/56/BRG 17 presented to the Scientific Committee of the International Whaling Commission, June 2004.

[24] J.B.S. Haldane. An exact test for randomness of mating. *Genetics*, 52:631–635, 1954.

[25] Kurt Hornik. The R FAQ. http://CRAN.R-project.org/doc/FAQ/R-FAQ.html, 2006. ISBN 3-900051-08-9.

[26] National Cancer Institude. Understanding Cancer Series. http://www.cancer.gov/cancertopics/UnderstandingCancer, 2005.

[27] J.E.Zeh and A.E.Punt. Updated 1978-2001 abundance estimates and their correlations for the Bering-Chukchi-Beaufort stock of bowhead whales. Paper IWC SC/56/BRG1 presented to the Scientific Committee of the International Whaling Commission, June 2004.

[28] T.Meeus F.Rousset J.Goudet, M.Raymond. Testing differentiation in diploid populations. *Genetic Society of America*, 144:1933–1940, 1996.

[29] J.K.Pritchard, M.Stephens, and P.Donnelly. Inference of population structure using multilocus genotype data. *Genetic Society of America*, 155:945–959, 2000.

[30] J.L.Roach, P.Stapp, B.V.Horne, and M.F.Antolin. Genetic structure of a metapopulation of black-tailed prairie dogs. *Journal of Mammalogy*, 82:946–959, 2001.

[31] K.E.W.Shelden and D.J.Rugh. The bowhead whale, balena mysticetus: Its historic and current status. *Marine Fisheries Review*, 57:1–20, 1995.

[32] K.J.Dawson and K.Bekhir. A bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genetical Research*, 78:59–77, 2001.

[33] K.L.Ayres and A.D.J.Overall. Allowing for within-subpopulation inbreeding in forensic match probabilities. *Forensic Science International*, 103:125–140, 1999.

[34] L.Excoffier. ARLEQUIN: A software for population genetics data analysis. http://anthro.unige.ch/arlequin, 2004.

[35] Mathsoft. *S-PLUS 2000 Guide to Statistics, Volume I.* Data Analysis Products Division, MathSoft, Seattle, WA, 1999.

[36] M.F.Antolin and W.C.Black. Gens, description of. *Encyclopedia of Biodiversity*, 3:183–193, 2001.

[37] M.Nei. F-statistics and analysis of gene diversity in subdivided populations. *Annals of Human Genetics*, 41:225–233, 1977.

[38] M.Nei and R.K.Chesser. Estimation of fixation indices and gene diversity. *Annals of Human Genetics*, 47:253–259, 1983.

[39] M.Raymond and F.Rousset. An exact test for population differentiation. *Evolution*, 49:1280–1283, 1995.

[40] M.Raymond and F.Rousset. Testing heterozygote excess and deficiency. *Genetics Society of America*, 140:1413–1419, 1995.

[41] N.A.Rosenberg. DISTRUCT: a program for the graphical display of population structure. *Moelcular Ecology Notes*, 4:137–138, 2004.

[42] P.E.Jorde, T.Schweder, and N.C.Stenseth. The Bering-Chukchi-Beaufort stock of bowhead whales: one homogeneous population? Paper IWC SC/56/BRG36 presented to the Scientific Committee of IWC, June 2004.

[43] D.S. Dittmer P.L.Altman. *Biology Data Book.* Federation of American Societies for Experimental Biology, 2nd edition, 1972-74.

[44] P.McCullagh and J.A.Nelder. *Generalized Linear Models.* Chapman & Hill, NY, 1989.

[45] P.W.Hendrick. *Genetics of Population.* Jones and Bartlett Publishers, 2nd edition, 2000.

[46] R.A.Fisher. *Statistical Methods for Research Workers.* Hafner, 13th edition, 1925.

[47] M. Raymond and F. Rousset. GENEPOP (ver. 1.2): A population genetics software for exact test and ecumenicism. *Journal of Heredity*, 86:248–249, 1995.

[48] M. Raymond and F. Rousset. GENEPOP (ver. 3.4): Software user manual. 2004.

[49] J.M.Olson R.C.Elston and L.Palmer. *Biostatistical Genetics and Epidemiology.* John Wiley and Sons, 2002.

[50] R.R.Sokal and F.J.Rohlf. *Biometry*, 1995.

[51] R.S.Waples and O.Gaggiotti. What is population? an emprical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Molecular Ecology*, 15:1419–1439, 2006.

[52] S.E.Moore and R.R.Reeves. Distribution and movement. 1999.

[53] S.James, A.Wylie, M.Johnson, S.Carstairs, and G.Simpson. Complex hybridity in isotoma pertaea. *Heredity*, 51:653–663, 1983.

[54] S.K.Wasser, A.M.Shedlock, K.Comstock, E.A.Ostrander, B.Mutayoba, and M.Stephens. Assigning African elephant DNA to geographic region of origin: applications to the ivory. *PNAS*, 101:14847–14852, 2004.

[55] S.N.Wood. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of American Statistical Association*, 99:673–686, 2004.

[56] S.Wahlund. Zusammensetzung von population und korrelationserscheinung vom standpunkt der vererbungslehre aus betrachtet. *Hereditas*, 106:11–65, 1928.

[57] S.W.Guo and E.A.Thompson. Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics*, 48:361–372, 1992.

[58] T.E.Maniatis, E.F.Fristch, and J.Sambrook. Molecular cloning: a laboratory manual., 1982.

[59] Laura A. Thompson. S-PLUS (and R) manual to accompany Agresti's categorical data analysis (2002). 2006.

[60] W.J.Richardson. Marine mammal and accustical monitoring of western geophysical's open-water seismic program in the alaska beaufort sea, 1998. 1999.

[61] S. Wright. *Evolution and the genetics of populations. Vol 2. The theory of gene frequencies.* University of Chicago Press, Chicago, 1969.

[62] F. Yates. Test of significance for 2x2 contingency tables. *Journal of Royal Statistics Society Assoc.*, 147:426–463, 1954.