UNIVERSITY OF BERGEN
Faculty of Mathematics and Natural Sciences

Statistical Approach to Relatedness Analysis in Large Collections of Genetic Profiles

an Application to a DNA-Registry of Fin Whales

Stefanía Benónísdóttir

M.S. - Thesis in
Statistics - Data Analyse
Supervisor: Professor Hans Julius Skaug
Co-supervisor: Dr. Christophe Pampoulie
November 2012

Contents

1 Introduction **1**

2 The Fin Whale **4**

3 Pedigree Analysis **6**
3.1 Mendelian Heritance Rules 6
3.2 Hardy-Weinberg and Linkage Equilibrium 7
3.3 Gene Identity by Descent 8

4 Statistical Methods **10**
4.1 The LOD Score 10
4.2 Controlling the Error Rate 15
4.2.1 p-Value 15
4.2.2 Multiple testing 16
4.3 Explanatory example 18
4.3.1 Method 18
4.3.2 Data 19
4.3.3 LOD Scores: 19
4.3.4 p-Value 20
4.3.5 Interpretation of the Result 22

5 Analysis of the Fin Whale Database **23**
5.1 Method 23
5.2 Data 25
5.3 Population Allele Frequencies 25
5.4 Application 25
5.4.1 Half-Siblings 25
5.4.2 Parent-Offspring 28
5.4.3 First Cousins 35
5.5 Results 40

6 Discussions **47**

Appendices **53**

A Kinship coefficients **54**
A.1 Siblings . . . 54
A.2 Half-Siblings . . . 55
A.3 First Cousins . . . 56

B Relatedness Likelihood Ratios **58**
B.1 Full Siblings Likelihood Ratio at a Single Locus . . . 58
B.2 Half-Siblings Likelihood Ratio at a Single Locus . . . 60
B.3 First Cousins Likelihood Ratio at a Single Locus . . . 61

C R Codes for the Explanatory Example **63**
C.1 Simulation of Individuals . . . 63
C.2 Computation of LOD Scores . . . 64
C.3 Estimation of p-Values . . . 69

D R Codes for the Fin Whale Analysis **72**
D.1 Registration of the Genetic Data . . . 72
D.2 Estimation of Population Allele Frequencies . . . 72
D.3 LOD Scores . . . 75
D.4 Simulation of Individuals . . . 78
D.5 Computation of p-Values . . . 78
D.6 Exact Binomial Confidence Interval . . . 79

E Estimated Population Allele Frequencies for the Fin Whale Analysis **80**

List of Tables

3.1 Kinship coefficients 9

4.1 Nr. of ways to inherit 0, 1 and 2 alleles IBD 11
4.2 Evaluation of multiple LOD scores 17
4.3 DNA profiles and allele frequencies for explanatory example 19
4.4 Pairwise LOD scores for parent-offspring hypothesis 20
4.5 Pairwise LOD scores for identical twins hypothesis 20
4.6 Pairwise LOD scores for siblings hypothesis 21
4.7 Pairwise LOD scores for half-siblings hypothesis 21
4.8 Pairwise LOD scores for first cousins hypothesis 22

5.1 50 highest pairwise half-sibling LOD scores and their corresponding $\hat{p}$-values . . 27
5.2 50 highest pairwise half-sibling LOD scores and their corresponding Bonferroni corrected $\hat{p}$-values 29
5.3 50 highest pairwise half-sibling LOD scores and their corresponding Q_r-values . 30
5.4 28 highest pairwise parent-offspring LOD scores and their corresponding $\hat{p}$-values. 32
5.5 28 highest pairwise parent-offspring LOD scores and their corresponding Bonferroni corrected $\hat{p}$-values. 33
5.6 28 highest pairwise parent-offspring LOD scores and their corresponding Q_r-values 34
5.7 9 highest pairwise parent-offspring LOD scores, mother-foetus pairs not included, and their corresponding Q_r-values 34
5.8 50 highest pairwise first cousins LOD scores and their corresponding $\hat{p}$-values . 36
5.9 50 highest pairwise first cousins LOD scores and their Bonferroni corrected $\hat{p}$-values 37
5.10 50 highest pairwise first cousins LOD scores and their corresponding Q_r values 38
5.11 33 highest pairwise first cousins LOD scores, mother-foetus pairs not included, and their corresponding Q_r values 39
5.12 Results from the FDR procedure with $q = 0.05$ not including non mother-foetus pairs 41
5.13 DNA profiles of F09-091, F09-091F and F10-100 46
5.14 Detected pairs of relatives within the sample 46

E.1 Allele frequencies at locus 1 to 6 80
E.2 Allele frequencies at locus 7 to 12 81
E.3 Allele frequencies at locus 13 to 15 81

Acknowledgements

I am very grateful to my supervisor Hans Julius Skaug for his great advices, for introducing me to statistical methods that I had not learned about before and for making sure that my study was always on track. Sincere thanks to Christophe Pampoulie, my co-supervisor in Iceland, for his contributions, support and for explaining genetics to me.

Many thanks to Bjarki Þór Elvarsson, who made it possible for me to do sufficiently many simulations with his technical support and gave me countless advice on formatting. Special thanks to Valérie Chosson.P., who answered any question I had about fin whales and provided estimation of the age and age of maturity of fin whales within the database. The people at the Marine Research Institute of Iceland are especially thanked for letting me work at their facilities and providing me with data.

Abstract

The use of DNA-profiles for identification is a matter of statistics. In the interpretation of genetic evidence there is always some uncertainty and this uncertainty requires estimation. The fin whale, *Balaenoptera physalus*, is a marine mammal that can be found in all of the world's oceans (Víkingsson, 2005). Fin whales, like all marine mammals, are by nature difficult to observe and uncertainties remain about their genetic structure, abundance, mating strategies and migration patterns (Pampoulie et al., 2012, Víkingsson, 2005, Ægisson and Hlíðberg, 2010). Introduction of DNA evidence at the end of the 1980s opened up many areas of research (Balding, 2005) but many relatedness studies based on genetic profiles have now been conducted for various species of wildlife (Nielsen et al., 2001, Skaug and Oien, 2005, Russell et al., 2009). This procedure has been especially useful for species that are difficult to observe because the identification of biological relationships yields information that can be useful in understanding the dynamics of species (Skaug et al., 2010, Pampoulie et al., 2012). Iceland has maintained an individual-based DNA-registry for fin whales for some time. The present study utilized data from this registry by searching for pairs of relatives among 267 fin whales and 23 fin whale foetuses. Three kind of relatedness were of interest, half-siblings, parent-offspring and first cousins. The LOD score is a commonly used test statistic for a given relatedness hypothesis and can be easily calculated for a pair of DNA profiles (Skaug et al., 2010). The LOD score is the logarithm of the ratio of the probabilities of the data under the two mutually exclusive hypotheses, H_0: *Unrelated* and $H1$: *Relatedness of interest*. Detection of relatives was done by computing pairwise LOD scores for the individuals in the sample for each relatedness of interest. The corresponding p-values for each LOD score were estimated by comparing the original LOD scores with LOD scores of unrelated individuals simulated with the same allele frequencies as the original dataset. Due to very high number of pairwise LOD scores it was necessary to adjust for multiple testing. Two well known multiple adjusting methods were applied and compared, the Bonferroni procedure and Benjamini's and Hochberg's (1995) false discovery rate procedure, (FDR). The FDR procedure was found to be more suitable for this analysis since the Bonferroni procedure was too conservative for such a high number of LOD scores. Eight pairs of relatives were detected within the sample at the false discovery rate of $q = 0.05$. When information about estimated age and estimated age of maturity had been taken into account, three of those pairs were classified as a parent and an offspring pair, two of them were classified as either half-siblings or an uncle/aunt-nephew/niece pair and the remaining three were classified as half-siblings or an uncle/aunt-nephew/niece pair or a grandparent-grandchild pair. One of the detected parent-offspring pairs were a male fin whale and a foetus which were also detected as a father and his offspring by Pampoulie et al. (2012).

Chapter 1

Introduction

In this present paper an Icelandic individual-based DNA registry of fin whales is utilized for identifying pairs of related individuals. The use of DNA-profiles for identification is a matter of statistics. In the interpretation of genetic evidence there is always uncertainty of some kind and this uncertainty requires estimation and statistical modelling. In their book 'Interpreting DNA Evidence: Statistical Genetics for Forensic Science', Evett and Weir (1998) suggested that genetic evidence should be interpreted according to three principles (Weir, 2007):

1. To evaluate the uncertainty for any given proposition it is necessary to consider at least one alternative proposition.

2. Interpretation is based on questions of the kind 'What is the probability of the evidence given the proposition?'

3. Interpretation is conditioned not only by the competing propositions, but also by the framework of circumstances within which they are to be evaluated.

The first principle leads to the use of likelihood ratios and the second one entails, in the present study, the question: 'Given that individual i and j are related, what is the probability that their DNA-profiles are as they are?' The third principle addresses the importance of taking auxiliary evidence into account. The first and second principle play a key role in the structure of the test procedure in the present analysis. Interpretation of the test results is conditioned by the non genetic evidence, as the third principle suggests, but estimation of age and age of maturity are used to conclude if parent-offspring and grandparent-grandchild relations are possible. Mother-foetus pairs are present within the DNA registry which is of great value for the analysis since the performance of the test procedure can be evaluated by it's ability to classify those mother-foetus pairs as relatives.

The basic idea in detecting related pairs of individuals based on genetic evidence, is that relatives share more alleles on average than unrelated individuals. To put it as simple as possible: 'The more a like the DNA profiles of individuals are, the higher probability that they are related'. It gets more complicated when relevant statistical genetic issues are incorporated into that probability. Population allele frequencies have to be accounted for. A match of two individuals that have DNA profiles that mainly contain common allele types does not have as much statistical power as a match of two individuals that have DNA profiles that consist of rare allele types. In this paper a match refers to finding a pair of individuals that the test

procedure concludes to be relatives. There are issues of how the population is to be defined and hence how the population allele frequencies should be estimated. There are also issues of assuming independence between loci and independence of segregation of alleles. These are all statistical genetic issues that will be addressed in chapter 3.

The LOD score is a well known test statistic for testing relatedness (Skaug et al., 2010). The LOD score is the logarithm, with base 10, of the likelihood ratio

$$LR = \frac{P(data \mid related)}{P(data \mid unrelated)}$$

The LOD score is used in the present analysis to test the hypothesis of a specific relatedness against the null hypothesis of unrelatedness and the data refers to DNA-profiles of a pair of fin whales. A p-value is estimated for each LOD score by simulating unrelated individuals from a population with the same allele frequencies as the observed sample, computing their pairwise LOD scores and comparing them to the LOD scores observed in the study.

According to the 2012 annual report of the Marine Institute in Iceland, 'Hagrannsóknir nr. 163', whale counting survey in 2007 indicated that there were about 20 600 fin whales in the East Greenland/Iceland tribe area (Sigurðsson and Magnússon, 2012). The database utilized in this study consists of genetic profiles of 267 fin whales and 23 foetuses from that area. That is a small fraction of the total population, $\frac{290}{20\ 600} \approx 0.0141$ which leads to a low probability of that both members of pairs of related individuals are within the sample.

A search for pair of relatives in a dataset of this size results in the simultaneous evaluation of $\frac{289 \cdot 290}{2} = 41\ 905$ pairwise LOD scores. This high number of test statistics entails a multiple comparison problem. If the significance level α would be used as in the single comparison case then the expected number of false detections would be considerable large or $m \cdot \alpha$, where m stands for the number of pairwise comparisons. In the present study the multiple comparison problem is accounted for by comparing two well known adjustment methods, the Bonferroni correction and the false discovery rate procedure. The Bonferroni correction controls the upper level of the family wise error rate, (FWER), which is defined as the probability of one or more type I error occurring in the analysis (Pounds et al., 2007). The Bonferroni correction entails testing each individual hypothesis at a level α/m which guarantees that $FWER \leq \alpha$. The Bonferroni procedure is very strict in the case of a high m with the cost of an increase in the number of type II errors[1], that is not detecting true relatives (Skaug et al., 2010). In 1995 Benjamini and Hochberg introduced the false discovery rate, (FDR), as a method to adjust for multiple testing. They described the FDR as an error rate that controls the expected proportion of false discoveries but in the present paper discoveries stands for detection of dyads of relatives. The FDR procedure arranges the estimated p-values for each LOD score in an increasing order $p_1 \leq p_2 \leq .... \leq p_m$. q is defined as the target false discovery rate and R is defined to be the largest value of r for which:

$$p_r \leq \frac{r}{m} \cdot q$$

The first R pairs of individuals that are behind the R lowest p-values in the ordered sequence are classified as relatives, and the remaining $m - R$ pairs are declared as unrelated. Benjamini

[1] Type II error is the error of not rejecting a false null hypothesis

and Hochberg (1995) showed in their paper that this procedure controls the false discovery rate at q for independent test statistics and for any configuration of false null hypotheses. The FDR is an appealing method to control the error rate in the present analysis of multiple pairwise LOD scores. It takes the number of erroneous false discoveries of relatedness into account instead of only the question whether any error was made (Benjamini and Hochberg, 1995). All computations and simulations in this study are done by using the open source program R, version 2.14.1, (R Development Core Team, 2011).

Chapter 2 contains information about fin whales. It is followed by the chapter 'Pedigree Analysis' which provides a brief introduction of the genetic concepts and theories needed for developing the statistical procedure used in the present study. The statistical procedure is presented in Chapter 4, 'Statistical Methods', and application of that procedure to the fin whale database is in the next chapter, 'Analysis of the Fin Whale Data Base'. In the last chapter 'Discussions', results are reviewed and conclusions drawn about the performance of the test procedure.

Chapter 2

The Fin Whale

The fin whale, *Balaenoptera physalus*, is a marine mammal that belongs to the suborder of baleen whales. Fin whales are considered to be the second longest animal in the world and can be as long as 25-27 meters (Víkingsson, 2005). The fin whale's body is greyish-black or brownish-black with a white underside. The dorsal fin is about 60 cm high and their spout is a direct 4-15 meters high pole. Fin whales usually don't dive for longer than 3-8 minutes and travel slowly but if they need to, fin whales can be underwater for 10-12 minutes and are capable of swimming as fast as 30 km per hour (Ægisson and Hlíðberg, 2010). Fin whales become mature when they are 7-12 years old (Víkingsson, 2005) and the oldest age record known is 114 years old (Ægisson and Hlíðberg, 2010). Fin whales travel alone or in small groups but whale counting in Icelandic and adjoining waters resulted in an average group size of 1.5 animal. (Víkingsson, 2005)

Fin whales can be found in all of the world's oceans. The International Whaling Commission has distinguished seven groups of fin whale in the North Atlantic but the fin whales around Iceland belong to the so called East Greenland/Iceland group, (EGI). This group distinction, which is largely based on distribution and development of whaling, occurrence during the summer and tagging, has been questioned but at the same time, no better well-founded hypothesis of division is available. In Icelandic waters, fin whales are most commonly seen during the summer and outside the tide land in the west and south-west of the country. The EGI group is considered to be the biggest fin whale group in the North Atlantic (Víkingsson, 2005). The Marine Research Institute of Iceland, in cooperation with neighbouring countries in the North Atlantic, has participated in wide-ranging whale counts in the years 1987, 1989, 1995, 2001 and 2007. According to those counts fin whales around Iceland have increased considerably in number since 1987, especially in the west of Iceland. Surveys from 1987-1989 indicated that there were about 16 000 fin whales in the EGI stock area. According to the survey in 2001 there were about 23 700 fin whales in all in the EGI stock area. The survey in 2007 indicated that 20 600 fin whales were in the EGI area. This estimate was not significantly different to that from 2001 but there is some uncertainty in a counting survey (Sigurðsson and Magnússon, 2012).

Fin whales, like all marine mammals, are by nature difficult to observe. It is mainly because they undertake long, annual migrations between high-latitude summer feeding areas and low-latitude winter breeding areas like most of the baleen whales. Little is known about the genetic

composition and biological characteristics of the group of individuals located at spawning grounds and available information have only been collected at feeding grounds. Therefore, despite that whales in Icelandic waters have been researched for centuries[1] uncertainties remain about fin whale's genetic structure, abundance, mating strategies and migration patterns (Ægisson and Hlíðberg, 2010, Víkingsson, 2005, Pampoulie et al., 2012).

The introduction of DNA techniques at the end of the 1980s opened up many areas of research (Balding, 2005). Genetic data can provide information on biological relationships between individuals (Skaug, 2001) but many relatedness studies based on genetic profiles have now been conducted for various species of wildlife (Nielsen et al., 2001, Skaug and Oien, 2005, Russell et al., 2009). This approach has been very useful for species that are difficult to observe, like fin whales, because the identification of biological relationships yields information that can be very useful in understanding the dynamics of species (Skaug et al., 2010, Pampoulie et al., 2012).

Iceland has maintained an individual-based DNA-registry for fin whales in recent years, which comprises 267 genetic profiles collected between and during the years 2009 and 2010, and has been obtained for 15 microsatellite loci, (neutral genetic markers inherited from the parents; see Pampoulie et al. (2012)), the control region of mtDNA and a sex-marker, (Bérubé and Palsbøll, 1996). 139 of the collected individuals were males. The database also contains information about the age and age of maturity of the individuals which was estimated by reading their ear plugs. 23 females, of the 267 individual samples genotyped, carried a foetus for which a genetic sample was also obtained (4 in 2009 and 19 in 2010). In the present project, available genetic profiles in the DNA registry of Iceland are examined to investigate if there are any biological relationships present between pairs of individuals within the database. Three relationships were of interest, half-siblings, parent-offspring and first cousins.

[1] In Konungs Skuggsjá, which was written in Norway in the 12th century, is an extensive description on whale species around Iceland. There is no doubt that the author got acquainted with whales in some way because some of the descriptions reconciles with what is best known about whales today (Sigurjónsson, 1993).

Chapter 3

Pedigree Analysis

This chapter provides a brief introduction of the genetic concepts and theories needed for developing the statistical procedure, used in the present study, for detecting relatives.

3.1 Mendelian Heritance Rules

Modern genetics began when Mendel published his First Law in 1866 on the basis of studies of pea plants. In his experiments he studied heritable traits in peas and postulated that discrete characters, which are now called genes, pass from parents to offspring. He suggested that each pea plant carries two genes that determine any given characteristic. One of the two genes is received from the male parent plant, the other from the female parent plant. In the formation of an offspring a random one of the two genes is passed on (that is segregates) from parent to offspring. Different offsprings of the same parent result from independent segregations. Mendel was able to explain his observations with this theory which is often called Mendel's First Law, the law of segregation. Mendel's First Law covers much of genetics since peas, like mammals, are diploid. That is, they carry genes in pairs which can therefore segregate in the way described (Thompson, 1986, Speed and Zhao, 2007).

The DNA in a cell is divided into chromosomes-substrings of the genetic material. In the formation of new cells it is the chromosomes that segregate, rather than individual genes (Thompson, 1986). Chromosomes other than the sex chromosomes are called autosomes. A diploid, which is the organism of interest here, has two complete sets of each autosome. The reproductive cell, sperm in males and egg in females, is called a gamete. Each gamete consists of a single version of the chromosome but a fusion of a male gamete and a female gamete forms a fertilized egg (Hartl and Jones, 1998).

A locus is a particular position on a chromosome. A diploid holds two alleles at each locus, one maternally inherited and the other one paternally inherited (Skaug, 2001). Generally, alleles are labelled according to their type. Consider a single locus and denote the number of different allelic types existing at that locus by K. The unordered pair of alleles carried by an individual is his genotype. In this case the possible genotypes are (a_i, a_j) with $i, j = 1, ...K$. Individuals with two copies of one allele are homozygous at that locus, but if the two alleles are different, then they are heterozygous. According to Mendelian segregation, a homozygous parent must pass on the only allelic type he/she carries to his/her offspring, while a heterozygous parent

passes on either one of his/her two allele types, each with probability 0.5. This is the basis of pedigree analysis (Thompson, 1986).

Mendel also considered two or more heritable traits together and carried out experiments to determine how traits in peas were inherited together. His observations, sometimes known as Mendel's Law of Independent Segregation, indicated that during a gamete formation, the segregation of one gene-pair is independent of the other gene-pairs. That is, when two gene-pairs (a, b) and (c, d) segregate, each gamete will be equally likely to have genotypes (a, c), (a, d), (b, c) and (b, d). Mendel's Law of Independent Segregation holds for some but not all pair of genes. It turns out that there are many pairs of traits whose genes do not recombine freely but tend to stick together, in the sense that parent with a genotypes (a, b) and (c, d) at two loci would be more likely to pass on the pairs (a, c) and (b, d) to his/her offspring than the pairs (a, d) and (b, c). This non independent segregation is known as linkage (Speed and Zhao, 2007).

3.2 Hardy-Weinberg and Linkage Equilibrium

Hartl and Jones (1998) define population as a group of organism of the same species living within a prescribed geographical area. This geographical area can be of any size but is commonly considered to be the area in which individuals within the population are likely to find mates (Hartl and Jones, 1998). The present study assumes that the fin whales with available genetic profiles at the DNA registry, all belong to the same population. However, there's no attempt to define the geographical area in which these individuals are likely to find mates.

The complete set of genetic information within a population is called a gene pool but the gene pool includes all alleles present in the population (Hartl and Jones, 1998). Allele types occur within different populations with different frequencies. A population allele frequency is the probability that a randomly chosen gene from a gene pool will be of a specific allelic type. That is, population allele frequencies give information on how common allele types are within the population (Thompson, 1986).

Consider S loci and denote by K_s the number of different allelic types that exist at locus s with $s = 1, ...S$. Assume the existence of a population with infinitely many individuals. The DNA-profile of individual i in the population is denoted by:

$$D_i = \{(a_{i,s}^{(1)}, a_{i,s}^{(2)}), 1 \leq s \leq S\} \tag{3.1}$$

$(a_{i,s}^{(1)}, a_{i,s}^{(2)})$ are unordered values. The population allele frequencies are denoted by $p(1_s), p(2_s)$, $..., p(K_s)$ with $\sum_{k=1}^{K_s} p(k_s) = 1$ but $p(k_s)$ is the population frequency for allele type k_s at locus s. The allele frequencies are obtained by dividing the observed number for each allele type by the total number of alleles in the gene pool (Hartl and Jones, 1998). If the genes can be regarded as independently chosen from an infinitely large gene pool with the above frequencies then the probability that the alleles at locus s of a randomly chosen individual i are of the same type is:

$$P((k_s, k_s)) = p(k_s)^2 \tag{3.2}$$

and the probability that the randomly chosen individual i has the genotype (k_s, r_s) at locus s with $k_s \neq r_s$ is:

$$P(k_s, r_s) = 2 \cdot p(k_s) \cdot p(r_s) \tag{3.3}$$

for $1 \leq i \leq \infty$ and $1 \leq s \leq S$. The genotype is an unordered pair, so the k_s allele may be chosen and then r_s or vice versa giving the factor of 2. These frequencies are known as the Hardy-Weinberg equilibrium frequencies (Thompson, 1986) but the population is said to be in Hardy Weinberg equilibrium if the alleles $a_{i,s}^{(1)}$ and $a_{i,s}^{(2)}$ are independent $\forall i$ (Balding, 2005). The population is said to be in linkage equilibrium if the genotypes $(a_{i,s}^{(1)}, a_{i,s}^{(2)})$ and $(a_{i,s'}^{(1)}, a_{i,s'}^{(2)})$ are independent for $s \neq s'$ and $\forall i$ (Skaug, 2001). Hardy-Weinberg and linkage equilibrium rarely hold in real populations since gene pools are never infinite. They can however provide a good approximation if the population size is large, mating is random and allele frequencies remain constant from one generation to the other (Balding, 2005, Hartl and Jones, 1998).

The present study is performed under the assumption of Hardy-Weinberg and linkage equilibrium. That entails the assumptions that the population is large enough, the mating is random and that major forces that influence allele frequencies, mutation, migration and selection, can be neglected. In random mating, organisms form mating pairs independently of genotype. Random mating is by far the most prevalent mating system for most species of animals (Hartl and Jones, 1998) and there is nothing that implies that mating among fin whales is an exception to that. Also, even if sexual selection was the case among fin whales, that would not necessary result in a Hardy-Weinberg disequilibrium. One important implication of the Hardy-Weinberg equilibrium is that the allele frequencies remain constant from one generation to the other (Hartl and Jones, 1998). Mutation is defined as a random change of the allelic type when an allele is passed from parent to offspring (Thompson, 1986). This change occurs with a very small probability but genes rarely undergo mutation in a single generation (Hartl and Jones, 1998). The generation time for fin whales is about 100 years. The genetic database that is utilized in this paper is from a lot shorter time span than one generation, 2009-2010, and therefore mutation is regarded as a negligible force. Selection is the differing viability and/or fertility of individuals according to their genotype. Since selection forces can be very complex and are seldom known with sufficient accuracy (Thompson, 1986) those forces will not be incorporated into the application. Migration of individuals, within or between populations, can have substantial effects over short periods (Balding, 2005, Thompson, 1986). The present study is the analysis of specified individuals so migration is, in this case, not relevant.

3.3 Gene Identity by Descent

The word *relatives* refers to individuals with common ancestors. Individuals will be found to have common ancestors if their ancestry is traced back far enough. For the purposes of this study, individuals are considered unrelated unless a precise relationship is specified. Every individual carries two alleles at each locus, one inherited from the mother and the other one inherited from the father. Any given set of individuals may carry the same allele types at a locus since there are many copies of an allele within a population. However, relatives are more likely to do so, for they may carry copies of a single gene inherited from one common ancestor. Genes that are copies of a single gene in a common ancestor are considered to be

identical by descent, (IBD). Such identical genes must be of the same allelic type while non IBD ones may or may not be. The basic idea is that genetic profiles of relatives are similar because they may carry IBD genes. Generally, closer relationships give higher probabilities for genes to be identical (Thompson, 1986).

One of the simplest probabilities of gene identity by descent is the classical kinship coefficient. The kinship coefficient, k_j, is defined as the probability that a pair of individuals has inherited j alleles at a locus identical by descent given a certain relatedness. $k_j = P(j - ibd \mid H_1)$ with $j = 0, 1, 2$. Table 3.1 contains values of the kinship coefficients for different stages of relatedness.

Table 3.1: Kinship coefficients

Hypothesis	k_0	k_1	k_2
Unrelated	1	0	0
Parent-offspring	0	1	0
Identical twins	0	0	1
Siblings	1/4	1/2	1/4
Half-siblings	1/2	1/2	0
First cousins	3/4	1/4	0

Unrelated individuals, by the definition in this study, don't have any common ancestors and therefore have inherited zero alleles identical by descent with probability 1. An offspring inherits one allele from its parent at each locus no matter what. By definition, identical twins inherit 2 alleles identical by descent at each locus with probability 1. It is a little more complicated to find the kinship coefficients for siblings, half-siblings and first cousins but the formulation for those coefficients can be found in appendix A. Full siblings are able to inherit 0, 1 or 2 alleles IBD at a locus but the probabilities differ. Half-siblings and first cousins are able to inherit 0 or 1 allele IBD under the assumption of no inbreeding. It is impossible to distinguish between a pair of half-siblings, grandparent-grandchild pair and uncle/aunt-nephew/niece pair from genetic evidence alone (Weir, 2007). For that reason the term 'half-siblings' refers here to all those relations unless noted otherwise.

Chapter 4

Statistical Methods

This chapter introduces the statistical procedure used for identifying pairs of close relatives within the collection of genetic profiles available at the fin whale DNA-registry.

4.1 The LOD Score

A commonly used test statistic for a given hypothesis about relatedness is the LOD score, which can be easily calculated from a pair of DNA-profiles (Skaug et al., 2010). Let D_i and D_j be the DNA profiles of individual i and j and consider the two mutually exclusive hypotheses:
H_0 : unrelated
H_1 : relatedness of interest
The LOD score is the logarithm of the ratio of the probabilities of the data under the two hypotheses:

$$LOD_{i,j} = log(\frac{P(D_i, D_j \mid H_1)}{P(D_i, D_j \mid H_0)}) \tag{4.1}$$

It does not matter whether it is H_1 in numerator and H_0 in denominator or vice versa, as long as it is clear which has been used. The LOD-value measures the probability of the data given H_1 relative to the probability of the data given H_0 (Balding, 2005). $LOD_{i,j} > c$ means that the data is more likely under H_1 than under H_0 where c is some predefined critical value (Skaug et al., 2010). The advantage of using LOD scores instead of just likelihood ratios is that the nature of the logarithm makes comparison of different relatedness hypothesis very simple. This is demonstrated in the explanatory example in section 4.3

In the present analysis Hardy-Weinberg and linkage equilibrium are assumed. Linkage equilibrium refers to the independence of inheriting alleles between loci (Skaug, 2001). This assumption enables extension of the formulation at a single locus to multi-loci formulation by the simple act of multiplication. Consider allele information of individuals i and j at S loci. Under linkage equilibrium it is possible to test their relatedness by computing the likelihood ratio for each locus separately and then multiply those ratios together and take the logarithm to attain the LOD score. If $LR_{i,j}(s) = \frac{P(D_{i,s}, D_{j,s}|H_1)}{P(D_{i,s}, D_{j,s}|H_0)}$ is the likelihood ratio at locus s then:

$$LOD_{i,j} = \log(LR_{i,j}(1) \cdot LR_{i,j}(2) \cdot ... \cdot LR_{i,j}(S)) \tag{4.2}$$

Under the assumption of Hardy-Weinberg and linkage equilibrium, the probability of two microsatellite based DNA-profiles, D_i and D_j, given the null hypothesis of unrelatedness, can be expressed as:

$$P(D_i, D_j \mid unrelated) = \prod_{s=1}^{S} p(a_{i,s}^{(1)}) \cdot p(a_{i,s}^{(2)}) \cdot p(a_{j,s}^{(1)}) \cdot p(a_{j,s}^{(2)}) \tag{4.3}$$

where $(a_{i,s}^{(1)}, a_{i,s}^{(2)})$ is the genotype of individual i at locus s, S is the number of independent markers and $p(a_{i,s}^{(m)})$ is the population frequency for whatever type allele $a_{i,s}^{(m)}$ is with $m = 1, 2$ (Skaug, 2001). When unrelated individuals have identical alleles then it is because there are many copies of the same allele in a population, not because they have common ancestors. In this case there are no random events that require conditioning. Unrelated individuals share 0 alleles identical by descent at each locus and there's only one way for not sharing any alleles at each locus.

The corresponding expression for $P(D_i, D_j \mid related)$ is more complicated and is formulated here by assuming Mendelian segregation. Since linkage equilibrium is assumed it is possible for convenience sake to put $S = 1$ and then extend the formulation for one locus to multi loci formulation by the simple act of multiplication. Let $D_i = (a_i^{(1)}, a_i^{(2)})$ and $D_j = (a_j^{(1)}, a_j^{(2)})$ be the genotypes of individual i and j. If two alleles are identical and have the same origin as well, that is are IBD, then: $a_i^{(m)} \equiv a_j^{(m)}$, $m = 1, 2$. If two alleles are not IBD (it doesn't rule out that they are identical though) then $a_i^{(m)} \not\equiv a_j^{(m)}$ with $m = 1, 2$. As before $(a_i^{(1)}, a_i^{(2)})$ are unordered values $\forall i$ and there's no way of knowing which allele is the mother-allele and which one is inherited from the father. Denote λ_s as the number of ways two individuals can inherit $s = 0, 1, 2$ alleles IBD at a locus regardless of their relatedness. Table 3.1 contains the values for lambda.

Table 4.1: Nr. of ways to inherit 0, 1 and 2 alleles IBD

λ_0	λ_1	λ_2
1	4	2

There's only one way for two individuals to have inherited zero alleles IBD.

$$a_i^{(1)} \not\equiv a_j^{(1)} \not\equiv a_i^{(2)} \not\equiv a_j^{(2)}$$

There are four different ways for two individuals to have inherited one allele IBD.

$$\begin{aligned} a_i^{(1)} &\equiv a_j^{(1)} \cap a_i^{(2)} \not\equiv a_j^{(2)} \\ a_i^{(1)} &\equiv a_j^{(2)} \cap a_i^{(2)} \not\equiv a_j^{(1)} \\ a_i^{(2)} &\equiv a_j^{(1)} \cap a_i^{(1)} \not\equiv a_j^{(2)} \\ a_i^{(2)} &\equiv a_j^{(2)} \cap a_i^{(1)} \not\equiv a_j^{(1)} \end{aligned}$$

There are two different ways for two individuals to have inherited two alleles IBD.

$$a_i^{(1)} \equiv a_j^{(1)} \cap a_i^{(2)} \equiv a_j^{(2)}$$
$$a_i^{(1)} \equiv a_j^{(2)} \cap a_i^{(2)} \equiv a_j^{(1)}$$

The general formula for $P(D_i, D_j \mid H_1)$ is:

$$\begin{aligned}
P(D_i, D_j \mid H_1) &= \frac{1}{\lambda_0} \cdot P(0 - IBD \mid H_1) \cdot p(a_i^{(1)}) \cdot p(a_i^{(2)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \\
&+ \frac{1}{\lambda_1} \cdot P(1 - IBD \mid H_1) \\
&\cdot (p(a_i^{(2)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(1)} = a_j^{(1)}) \\
&+ (p(a_i^{(1)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(2)} = a_j^{(1)}) \\
&+ (p(a_i^{(2)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(1)} = a_j^{(2)}) \\
&+ (p(a_i^{(1)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(2)} = a_j^{(2)})) \\
&+ \frac{1}{\lambda_2} \cdot P(2 - IBD \mid H_1) \\
&\cdot (p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(1)} = a_j^{(1)} \bigcap a_i^{(2)} = a_j^{(2)}) \\
&+ p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(1)} = a_j^{(2)} \bigcap a_i^{(2)} = a_j^{(1)})) \\
&= \frac{1}{1} \cdot k_0(H_1) \cdot p(a_i^{(1)}) \cdot p(a_i^{(2)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \\
&+ \frac{1}{4} \cdot k_1(H_1) \\
&\cdot (p(a_i^{(2)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(1)} = a_j^{(1)}) \\
&+ (p(a_i^{(1)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(2)} = a_j^{(1)}) \\
&+ (p(a_i^{(2)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(1)} = a_j^{(2)}) \\
&+ (p(a_i^{(1)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(2)} = a_j^{(2)})) \\
&+ \frac{1}{2} \cdot k_2(H_1) \\
&\cdot (p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(1)} = a_j^{(1)} \bigcap a_i^{(2)} = a_j^{(2)}) \\
&+ p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(1)} = a_j^{(2)} \bigcap a_i^{(2)} = a_j^{(1)}))
\end{aligned} \tag{4.4}$$

When relatedness is determined from genetic profiles, without any auxiliary data, it is impossible to know if identical alleles are identical because they stem from the same origin or if they are just identical by chance. The kinship coefficients, $k_j(H_1) = P(j - IBD \mid H_1)$, incorporate that uncertainty into the formulation but they refer to the probability of two individuals inheriting j alleles IBD given the hypothesis of relatedness, $j = 0, 1, 2$. $p(a_i^{(1)})$ is the population allele frequency for whatever allele type $a_i^{(1)}$ is, $p(a_i^{(2)})$ is the population allele frequency for whatever allele type $a_i^{(2)}$ is etc. If $a_i^{(1)} = a_j^{(1)}$ then those alleles are of the same type and consequently $p(a_i^{(1)}) = p(a_j^{(1)})$. $I(a = b)$ is the identity function, that is $I(a = b) = 1$ if $a = b$ and $I(a = b) = 0$ otherwise. If this formulation is applied to the hypothesis of unrelatedness then that will result in formula 4.3 with $S = 1$:

$$
\begin{aligned}
P(D_i, D_j \mid H_0) = & \frac{1}{1} \cdot 1 \cdot p(a_i^{(1)}) \cdot p(a_i^{(2)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \\
& + \frac{1}{4} \cdot 0 \\
& \cdot (p(a_i^{(2)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(1)} = a_j^{(1)}) \\
& + (p(a_i^{(1)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(2)} = a_j^{(1)}) \\
& + (p(a_i^{(2)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(1)} = a_j^{(2)}) \\
& + (p(a_i^{(1)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(2)} = a_j^{(2)})) \\
& + \frac{1}{2} \cdot 0 \\
& \cdot (p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(1)} = a_j^{(1)} \bigcap a_i^{(2)} = a_j^{(2)}) \\
& + p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(1)} = a_j^{(2)} \bigcap a_i^{(2)} = a_j^{(1)})) \\
& = p(a_i^{(1)}) \cdot p(a_i^{(2)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)})
\end{aligned}
\tag{4.5}
$$

The computation of the LOD score can be time-consuming. Balding (2005) presents a simple formulation for the relatedness likelihood ratio on page 126:

$$
LR_{H_1} = k_0(H_1) + k_1(H_1) \cdot LR_p + k_2(H_1) \cdot LR_{id} \tag{4.6}
$$

LR_p : Likelihood ratio for parent-offspring relations.
LR_{id} : Likelihood ratio for identical twins relations.
k_0, k_1 and k_2 are relevant kinship coefficients.

According to this formula, to test diverse hypothesis of relatedness against the null hypothesis of unrelatedness, only the parent-offspring likelihood ratio, (LR_p), and the identity likelihood ratio, (LR_{id}), have to be computed. Below is the formulation of LR_p and LR_{id} for a single locus. Having those quantities fixed, various relations can be tested by using the appropriate kinship coefficients which can be found in table 3.1.

Likelihood ratio for parent-offspring relations compares the mutually exclusive hypothesis:
H_0 : Individual i and j are unrelated
H_1 : Individual i and j are a parent and his/her offspring.
There is one random event that needs conditioning, which allele was inherited from the mother and which allele was inherited from the father.

$$
\begin{aligned}
P(D_i, D_j \mid parent\ and\ offspring) = 0 & \\
& + \frac{1}{4} \cdot 1 \\
& \cdot (p(a_i^{(2)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(1)} = a_j^{(1)}) \\
& + p(a_i^{(1)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(2)} = a_j^{(1)}) \\
& + p(a_i^{(2)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(1)} = a_j^{(2)}) \\
& + p(a_i^{(1)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(2)} = a_j^{(2)})) \\
& + 0
\end{aligned} \tag{4.7}
$$

The parent-offspring likelihood ratio has formula 4.5 in the denominator and formula 4.7 in the numerator:

$$
\begin{aligned}
LR_p &= \frac{P(D_i, D_j \mid parent\ and\ offspring)}{P(D_i, D_j \mid not\ related)} \\
&= \frac{I(a_i^{(1)} = a_j^{(1)}) + I(a_i^{(1)} = a_j^{(2)})}{4 \cdot p(a_i^{(1)})} + \frac{I(a_i^{(2)} = a_j^{(1)}) + I(a_i^{(2)} = a_j^{(2)})}{4 \cdot p(a_i^{(2)})}
\end{aligned} \tag{4.8}
$$

The identity likelihood ratio compares the two mutually exclusive hypothesis:
H_0 : Individual i and j are unrelated.
H_1 : Individual i and j are identical twins
Here, there is no random event that needs conditioning on.

$$
\begin{aligned}
P(D_i, D_j \mid identical\ twins) = 0 + 0 + \frac{1}{2} \cdot 1 & \\
& \cdot (p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot (I(a_i^{(1)} = a_j^{(1)}) \bigcap I(a_i^{(2)} = a_j^{(2)})) \\
& + p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot (I(a_i^{(1)} = a_j^{(2)}) \bigcap I(a_i^{(2)} = a_j^{(1)})))
\end{aligned} \tag{4.9}
$$

The identity likelihood ratio is attained by incorporating formulas 4.5 and 4.9 into the likelihood ratio:

$$
\begin{aligned}
LR_{id} &= \frac{P(D_i, D_j \mid identical\ twins)}{P(D_i, D_j \mid not\ related)} \\
&= \frac{(I(a_i^{(1)} = a_j^{(1)}) \bigcap I(a_i^{(2)} = a_j^{(2)})) + (I(a_i^{(1)} = a_j^{(2)}) \bigcap I(a_i^{(2)} = a_j^{(1)}))}{2 \cdot p(a_i^{(1)}) \cdot p(a_i^{(2)})}
\end{aligned} \tag{4.10}
$$

Now it is rather basic to calculate the likelihood ratio for other hypothesis of relatedness.

$$
\begin{aligned}
LR_{sib} &= \frac{P(D_i, D_j \mid siblings)}{P(D_i, D_j \mid not\ related)} \\
&= \frac{1}{4} + \frac{1}{2} \cdot (\frac{I(a_i^{(1)} = a_j^{(1)}) + I(a_i^{(1)} = a_j^{(2)})}{4 \cdot p(a_i^{(1)})} + \frac{I(a_i^{(2)} = a_j^{(1)}) + I(a_i^{(2)} = a_j^{(2)})}{4 \cdot p(a_i^{(2)})}) \\
&+ \frac{1}{4} \cdot (\frac{(I(a_i^{(1)} = a_j^{(1)}) \bigcap I(a_i^{(2)} = a_j^{(2)})) + (I(a_i^{(1)} = a_j^{(2)}) \bigcap I(a_i^{(2)} = a_j^{(1)}))}{2 \cdot p(a_i^{(1)}) \cdot p(a_i^{(2)})})
\end{aligned} \tag{4.11}
$$

$$
\begin{aligned}
LR_{h.sib} &= \frac{P(D_i, D_j \mid half - siblings)}{P(D_i, D_j \mid not\ related)} \\
&= \frac{1}{2} + \frac{1}{2} \cdot \left(\frac{I(a_i^{(1)} = a_j^{(1)}) + I(a_i^{(1)} = a_j^{(2)})}{4 \cdot p(a_i^{(1)})}\right. \\
&\left. + \frac{I(a_i^{(2)} = a_j^{(1)}) + I(a_i^{(2)} = a_j^{(2)})}{4 \cdot p(a_i^{(2)})}\right)
\end{aligned} \tag{4.12}
$$

$$
\begin{aligned}
LR_{cous} &= \frac{P(D_i, D_j \mid first\ cousins)}{P(D_i, D_j \mid not\ related)} \\
&= \frac{3}{4} + \frac{1}{4} \cdot \left(\frac{I(a_i^{(1)} = a_j^{(1)}) + I(a_i^{(1)} = a_j^{(2)})}{4 \cdot p(a_i^{(1)})}\right. \\
&\left. + \frac{I(a_i^{(2)} = a_j^{(1)}) + I(a_i^{(2)} = a_j^{(2)})}{4 \cdot p(a_i^{(2)})}\right)
\end{aligned} \tag{4.13}
$$

Formulation that shows why it is sufficient to use the simple formula from Balding (2005) for the three likelihood ratios, given by equations 4.11, 4.12 and 4.13, can be found in appendix B.

4.2 Controlling the Error Rate

$LOD_{i,j} > c$ means that the data is more likely under H_1 : *Relatedness of interest* than under H_0: *Unrelated*, where c is some predefined critical value. Determining the value of c involves finding a balance between two goals, maximization of the number of correctly inferred pairs of relatives and minimization of the number of pairs incorrectly concluded as relatives(Skaug et al., 2010).

4.2.1 p-Value

Two types of error can occur in statistical hypothesis testing. A Type I error occurs when a true null hypothesis is rejected. A type II error is the error of not rejecting a false null hypothesis. The significance level of a test, denoted by α, is an upper bound on the probability of a Type I error. That means that if the test procedure would be replicated a large number of times under the conditions of H_0 then the observed Type I error rate should be at most α. An observed value of the test statistic is said to be significant if the test decision based on that statistic is to reject the null hypothesis. A p-value is defined as the smallest possible value of α such that the observed test statistic would be significant (Rizzo, 2007).

When testing a single pair of hypotheses one usually reports a single p-value with the test statistic. In the present study, a large value of the LOD score clearly provides evidence against unrelatedness but then there's an issue of what should be considered large enough for the LOD score to be significant. Evaluation of a p-value addresses that issue. For any given test statistic

there are often several ways to compute a p-value (Pounds et al., 2007). Here, the p-values are evaluated with a Monte Carlo experiment built on a permutation method that Skaug et al. (2010) performed in the analysis of minke whale data. Let $p_{i,j}$ be the corresponding p-value for $LOD_{i,j}$. $p_{i,j}$ is estimated in the following steps:

Step 1. r unrelated individuals are simulated with the allele frequencies that were estimated from the original data set.

Step 2. The $m = \frac{r(r-1)}{2}$ pairwise LOD scores of the simulated individuals are computed.

Step 3. The test decisions are recorded:

$$\begin{aligned} I_w^{i,j} &= 1, \quad LOD_w \geq LOD_{i,j} \\ I_w^{i,j} &= 0, \quad LOD_w < LOD_{i,j} \end{aligned} \tag{4.14}$$

LOD_w stands for the simulated LOD score with $w = 1, 2, ..., m$.

Step 4. The proportion of simulated LOD-scores that are equal or larger than the original test statistic $LOD_{i,j}$ is calculated to attain the $\hat{p}$-value.

$$\hat{p}_{i,j} = \frac{\Sigma_{w=1}^{m} I_w^{i,j}}{m} \tag{4.15}$$

Step 5. The procedure is repeated until the standard errors of the $\hat{p}$-values are sufficiently small. The $\hat{p}$-values are binomial distributed so if M is the total number of simulated LOD scores then the standard deviation of $\hat{p}_{i,j}$ is:

$$sd(\hat{p}_{i,j}) = \sqrt{\frac{\hat{p}_{i,j} \cdot (1 - \hat{p}_{i,j})}{M}} \tag{4.16}$$

In the present analysis it is the number of simulated LOD scores that matters, so the selection of r, the number individuals that are simulated each time, simply depends on what is the most convenient way to attain sufficiently large M in the end. In principle, the above simulation procedure must be repeated for every single pair of individuals within the original dataset. However, to reduce computational burden, the set of M simulated LOD scores are kept fixed across all pairwise comparisons (Skaug et al., 2010, Rizzo, 2007).

4.2.2 Multiple testing

A decision to reject the null hypothesis in a single comparison is usually made by comparing the p-value to the customary significance level. If the significance level is fixed at $\alpha = 0.05$, on average one in every 20 LOD scores of unrelated individuals will show a p-value below α just by chance. If the significance level α would be used as in the single comparison case then the expected number of false detections in the analysis would be $m \cdot \alpha$, where m stands for the number of pairwise comparisons. Due to the large number of comparisons in the present analysis this expected number of false detections is considerable large, especially when the low probability of detection, due too how small the sample is compared to the estimated population size, is taken into account. For this reason it is important to adjust the estimated p-values for multiple testing (Skaug et al., 2010, Casella and Berger, 2002, Johnson and Wichern, 2007).

Bonferroni

Traditional approaches to adjust for multiple testing attempt to control the family wise error rate. The family wise error rate is defined as the probability that one or more Type I error occur in the group of hypothesis tests (Pounds et al., 2007). Benjamini and Hochberg (1995) denoted the family wise error rate by $FWER = P(V \geq 1)$, where V stands for the number of falsely rejected null hypothesis. They pointed out that when m hypothesis are tested individually at a level α guarantees that $E[V/m] \leq \alpha$. The Bonferroni procedure entails testing each individual hypothesis at a level α/m which guarantees that $P(V \geq 1) \leq \alpha$. In the present study the Bonferroni adjustment is done by multiplying the unadjusted p-values by the number of pairwise comparisons and then compare them with α (Huber et al., 2007). The Bonferroni procedure is usually considered too conservative in the case of high numbers of test statistics but it is almost impossible to have any rejections of the null hypothesis with the Bonferroni correction in the context of thousand pairwise comparisons (Pounds et al., 2007). Since the present analysis involves a high number of multiple pairwise comparisons, the Bonferroni correction is probably to strict at the cost of not detecting true relatives.

The False Discovery Rate

In 1995 Benjamini and Hochberg introduced the False Discovery Rate, (FDR), as a method to adjust for multiple testing. They described the FDR as an error rate that controls the expected proportion of false discoveries. Consider the problem of evaluating simultaneously m LOD scores of which m_0 consist of pairs of unrelated individuals. R is the number of LOD-scores that result in rejection of the null hypothesis of unrelatedness.

Table 4.2: Evaluation of multiple LOD scores

	Declared unrelated	**Declared related**	**Total**
Truly unrelated	U	V	m_0
Truly related	T	S	$m - m_0$
Total	$m - R$	R	m

R is an observable random variable, U, V, S and T are unobservable random variables. If each individual hypothesis pair is tested separately at level α, then $R = R(\alpha)$ is increasing in α. The FDR is denoted by:

$$FDR = E[\frac{V}{V+S}] = E[\frac{V}{R}] \tag{4.17}$$

That is, the false discovery rate is the expectation of the random variable $\frac{V}{R}$. The FDR is an appealing method to control the error rate in the present analysis, with a very high number of pairwise tests, since it takes the number of erroneous false discoveries of relatedness into account instead of only the question whether any error was made.

The FDR procedure arranges the estimated p-values for each LOD score in an increasing order $\hat{p}_{(1)} \leq \hat{p}_{(2)} \leq .... \leq \hat{p}_{(m)}$. q is the target false discovery rate and R is defined to be the largest value of r for which:

$$\hat{p}_{(r)} \leq \frac{r}{m} \cdot q \tag{4.18}$$

The first R pairs of individuals that are behind the R lowest $\hat{p}$-values in the ordered sequence are declared as being related, and the remaining $m - R$ pairs are declared as unrelated. Benjamini and Hocberg (1995) showed in their paper that this procedure controls the false discovery rate at q for independent test statistics and for any configuration of false null hypotheses.

In the present analysis not all of the LOD scores are independent of each other as the Benjamini-Hochberg FDR procedure assumes. In a dataset of n individuals, each individual must be involved in $(n-1)$ pairwise tests and those $(n-1)$ LOD scores are not independent of each other. Skaug et al. (2010) pointed out that given two individuals i or j, the proportion of pairwise comparisons (among $m = n(n-1)/2$) involving either i or j is approximately $\frac{4}{n}$ and thus becomes negligible when n gets large. They conducted a Monte Carlo simulation to investigate how large n has to be for this result to apply. In their case, $n = 100$ seemed to be large enough for the FDR procedure to behave as expected, even if some pairwise comparisons were not independent.

4.3 Explanatory example

In this section relations between three simulated individuals, A, B and C will be investigated for explanatory purposes. The individuals were simulated with given allele frequencies so that B and C would be siblings and A would be unrelated to B and C. This simple example should demonstrate how the test procedure operates. All computations and simulations were done by using the open source program *R* (R Development Core Team, 2011) but the codes written can be found in appendix C.

4.3.1 Method

If one goes back far enough in the family tree, possible relations would be endless. In this example the following set of relatedness hypothesis will be tested against the null hypothesis of unrelatedness:

1. H_1: Individual i and individual j have a parent-offspring relationship, $i \neq j$, $i, j = A, B, C$.
2. H_1: Individual i and individual j are identical twins, $i \neq j$, $i, j = A, B, C$.
3. H_1: Individual i and individual j are siblings, $i \neq j$, $i, j = A, B, C$.
4. H_1: Individual i and individual j are half-siblings, $i \neq j$, $i, j = A, B, C$.
5. H_1: Individual i and individual j are first cousins, $i \neq j$, $i, j = A, B, C$.

A two step procedure for each relatedness hypothesis is applied to the data. The appropriate pairwise LOD-scores are computed, $LOD_{i,j}$ for $i \neq j$ in the first step. In the second step p-values for positive LOD scores are estimated. This is done by simulating 100 unrelated individuals with the same allele frequencies that individual A, B and C were simulated with.

This procedure is replicated 50 times[1] resulting in a total of 247 500 simulated LOD scores. The $\hat{p}$-value, $\hat{p}_{i,j}$ is the proportion of simulated LOD scores that are equal or higher than $LOD_{i,j}$. The only information available in this example are the genetic profiles. Recall that it is impossible to distinguish between a pair of half siblings, a pair of grandparent and a grand child and a pair of uncle/aunt and a nephew/niece from genetic evidence alone so the term 'half-siblings' refers to all those relations here.

4.3.2 Data

The genetic profiles of three individuals are simulated with frequencies as given in table 3.3. Individual B and C were simulated to be siblings while individual A was simulated to be unrelated to them. The data consists of information about 10 loci which are segregated by two alleles. The number of possible alleles varies between loci and the alleles are not ordered values, that is: $a/b = b/a$. The population allele frequencies are fixed for each locus and were computed by dividing 1 with the number of possible allele types at that specific locus.

Table 4.3: DNA profiles and allele frequencies for explanatory example

Locus	A	B	C	Allele frequencies
1	4/8	8/2	3/2	0.111
2	15/6	12/15	5/15	0.056
3	8/3	17/11	17/14	0.050
4	3/13	4/6	4/3	0.071
5	7/10	8/3	6/3	0.100
6	8/1	8/8	8/10	0.083
7	2/5	11/1	9/8	0.083
8	3/9	10/9	10/8	0.091
9	6/13	10/2	14/10	0.059
10	12/10	13/16	13/9	0.063

4.3.3 LOD Scores:

The likelihood ratios for each locus given the DNA data were computed in R but the LOD score is the logarithm of the multi loci likelihood ratio. The resulting likelihood ratios and LOD scores can be found in tables 3.4, 3.5, 3.6, 3.7 and 3.8. All the numbers have been rounded to numbers with two decimal places:

Of the twelve computed pairwise LOD scores, four of them are larger than zero. For individual B and C there are three non negative LOD scores: $LOD_{B,C}(sib) = 2.31$, $LOD_{B,C}(h.sib) = 3.06$, $LOD_{B,C}(cous) = 1.93$. For individual A and B there is one non negative LOD score: $LOD_{A,B}(cous) = 0.34$. All the other LOD scores don't provide evidence against unrelatedness since $LOD_{i,j} < 0$ means that $P(D_i, D_j \mid unrelated) \geq P(D_i, D_j \mid related)$.

[1] Here, for simplification, there is no concern for the standard deviation of the $\hat{p}$-values.

Table 4.4: Pairwise LOD scores for parent-offspring hypothesis

Parent-offspring	**A - B**	**A - C**	**B - C**
Locus 1: $LR_1(p)$	2.25	0.00	2.25
Locus 2: $LR_2(p)$	4.50	4.50	4.50
Locus 3: $LR_3(p)$	0.00	0.00	5.00
Locus 4: $LR_4(p)$	0.00	3.50	3.50
Locus 5: $LR_5(p)$	0.00	0.00	2.50
Locus 6: $LR_6(p)$	6.00	3.00	6.00
Locus 7: $LR_7(p)$	0.00	0.00	0.00
Locus 8: $LR_8(p)$	2.75	0.00	2.75
Locus 9: $LR_9(p)$	0.00	0.00	4.25
Locus 10: $LR_{10}(p)$	0.00	0.00	4.00
$LOD(p) = log(\prod_{s=1}^{10} LR_s(p))$	$-\infty$	$-\infty$	$-\infty$

Table 4.5: Pairwise LOD scores for identical twins hypothesis

Identical twins	**A - B**	**A - C**	**B - C**
Locus 1: $LR_1(id)$	0.00	0.00	0.00
Locus 2: $LR_2(id)$	0.00	0.00	0.00
Locus 3: $LR_3(id)$	0.00	0.00	0.00
Locus 4: $LR_4(id)$	0.00	0.00	0.00
Locus 5: $LR_5(id)$	0.00	0.00	0.00
Locus 6: $LR_6(id)$	0.00	0.00	0.00
Locus 7: $LR_7(id)$	0.00	0.00	0.00
Locus 8: $LR_8(id)$	0.00	0.00	0.00
Locus 9: $LR_9(id)$	0.00	0.00	0.00
Locus 10: $LR_{10}(id)$	0.00	0.00	0.00
$LOD(id) = log(\prod_{s=1}^{10} LR_s(id))$	$-\infty$	$-\infty$	$-\infty$

4.3.4 p-Value

p-values were estimated for the non negative LOD-scores by simulation. First 100 unrelated individuals were simulated from a population with the same allele frequencies that individual A, B and C were simulated with. Then their 4 950 pairwise LOD-scores are computed. This procedure is replicated 50 times resulting in 247 500 simulated LOD scores for each relatedness hypothesis. The $\hat{p}$-value is the proportion of simulated LOD scores that are equal or higher than the original LOD-score. The following $\hat{p}$-values were attained:

Table 4.6: Pairwise LOD scores for siblings hypothesis

Siblings	**A - B**	**A - C**	**B - C**
Locus 1: $LR_1(sib)$	1.38	0.25	1.38
Locus 2: $LR_2(sib)$	2.50	2.50	2.50
Locus 3: $LR_3(sib)$	0.25	0.25	2.75
Locus 4: $LR_4(sib)$	0.25	2.00	2.00
Locus 5: $LR_5(sib)$	0.25	0.25	1.50
Locus 6: $LR_6(sib)$	3.25	1.75	3.25
Locus 7: $LR_7(sib)$	0.25	0.25	0.25
Locus 8: $LR_8(sib)$	1.63	0.25	1.63
Locus 9: $LR_9(sib)$	0.25	0.25	2.38
Locus 10: $LR_{10}(sib)$	0.25	0.25	2.250
$LOD(sib) = log(\prod_{s=1}^{10} LOD_s(sib))$	-2.35	-3.27	2.30

Table 4.7: Pairwise LOD scores for half-siblings hypothesis

Half-siblings	**A - B**	**A - C**	**B - C**
Locus 1: $LR_1(h.sib)$	1.63	0.50	1.63
Locus 2: $LR_2(h.sib)$	2.75	2.75	2.75
Locus 3: $LR_3(h.sib)$	0.50	0.50	3.00
Locus 4: $LR_4(h.sib)$	0.50	2.25	2.25
Locus 5: $LR_5(h.sib)$	0.50	0.50	1.75
Locus 6: $LR_6(h.sib)$	3.50	2.00	3.50
Locus 7: $LR_7(h.sib)$	0.50	0.50	0.50
Locus 8: $LR_8(h.sib)$	1.88	0.50	1.88
Locus 9: $LR_9(h.sib)$	0.50	0.50	2.63
Locus 10: $LR_{10}(h.sib)$	0.50	0.50	2.50
$LOD(h.sib) = log(\prod_{s=1}^{10} LR_s(h.sib))$	-0.34	-1.01	3.06

$$p_{B,C}(sib) = \frac{10\ 414}{247\ 500} = 4.21 \cdot 10^{-2}$$
$$p_{B,C}(h.sib) = \frac{4}{247\ 500} = 1.62 \cdot 10^{-5}$$
$$p_{B,C}(cous) = \frac{8}{247\ 500} = 3.23 \cdot 10^{-5}$$
$$p_{A,B}(cous) = \frac{24\ 340}{247\ 500} = 9.83 \cdot 10^{-2}$$

In this example there are only three pairwise comparisons for each relatedness hypothesis so there's no need for multiple comparison adjustment. H_1 : *Individual A and individual B are first cousins* is rejected at the significance level of $\alpha = 0.05$ while the relatedness hypothesis for individual B and C can not be rejected at that significance level. It is clear that the test procedure classifies individual B and C as relatives but in order to conclude about a specific relatedness, their three different LOD scores have to be compared. $LOD_{h.sib}$, LOD_p and LOD_{cous} compare the probability of the data under the hypothesis of a specific

Table 4.8: Pairwise LOD scores for first cousins hypothesis

First cousins	**A - B**	**A - B**	**B - C**
Locus 1: $LR_1(cous)$	1.31	0.75	1.31
Locus 2: $LR_2(cous)$	1.88	1.88	1.88
Locus 3: $LR_3(cous)$	0.75	0.75	2.00
Locus 4: $LR_4(cous)$	0.75	1.63	1.63
Locus 5: $LR_5(cous)$	0.75	0.75	1.38
Locus 6: $LR_6(cous)$	2.25	1.50	2.25
Locus 7: $LR_7(cous)$	0.75	0.75	0.750
Locus 8: $LR_8(cous)$	1.44	0.75	1.44
Locus 9: $LR_9(cous)$	0.75	0.75	1.81
Locus 10: $LR_{10}(cous)$	0.75	0.75	1.75
$LOD(cous) = log(\prod_{s=1}^{10} LR_s(cous))$	0.34	-0.21	1.93

relatedness with the probability of the data under the null hypothesis of unrelatedness. What is needed now is to compare the probability of the data under one specific relatedness with the probability of the data under another specific relatedness. That can be done by simply subtracting one LOD score from the other since $log(\frac{a}{b}) = log(a) - log(b) = log(\frac{a}{c}) - log(\frac{b}{c})$.

$$log(\frac{P(data \mid half - siblings)}{P(data \mid siblings}) = LOD_{h.sib} - LOD_{sib}$$
$$= 3.06 - 2.31 = 0.75$$

$$log(\frac{P(data \mid half - siblings)}{P(data \mid first\ cousins}) = LOD_{h.sib} - LOD_{cous}$$
$$= 3.06 - 1.93 = 1.13$$

The LOD scores indicate that individual B and C are more likely to be half-siblings than siblings or first cousins.

4.3.5 Interpretation of the Result

The procedure used in the present study for testing relatedness concluded siblings wrongly to be half-sibling but concluded an unrelated individual rightly as a non-relative. The fact that the procedure did not detect the right relatedness between individual B and C can be explained by the fact that they were simulated, by chance, as siblings that didn't have an identical genotype at any locus. The probability for full siblings to not have a single identical genotype at 10 loci is small or: $(\frac{3}{4})^{10} = 0.0563$. Therefore it is logical, though not correct in this case, that the test procedure indicated that the individuals were half-siblings rather than full siblings.

Chapter 5

Analysis of the Fin Whale Database

The goal of this analysis is to detect pairs of relatives within the Icelandic fin whale registry. Three relations are of interest, half-siblings, parent-offspring and first cousins.

5.1 Method

A three-step procedure for each relatedness hypothesis is applied to the fin whale data:

1. Pairwise LOD scores computed.
2. p-value for each LOD score estimated via simulation.
3. $\hat{p}$-values adjusted for multiple testing.

In the first step the appropriate pairwise LOD scores are computed from the dataset. A LOD score is a commonly used test statistic to detect related individuals within a database (Skaug et al., 2010) but it compares the probabilities of the data under the null hypothesis of unrelatedness and the alternative hypothesis of relatedness.

$$LOD = log(\frac{P(data \mid relatedness\ of\ interest)}{P(data \mid unrelated)})$$

log stands for the $10th$ logarithm. Further information about LOD scores can be found in chapter 4.1.

In the second step the corresponding p-values for each LOD score are estimated. There is a negative relationship between a p-value and its LOD score, $\frac{\delta p_{i,j}}{\delta LOD_{i,j}} < 0$, but a high LOD score and a low p-value indicate relatedness. To reduce computational burden, p-values are at first only estimated for the 1000 largest LOD scores. If the estimated p-value for the 1000th highest LOD score is high enough for the LOD score to be considered unsignificant then the conclusion is that all the lower LOD scores are unsignificant as well and further estimation is unnecessary. If the p-value for the 1000th highest LOD score is low enough for the LOD score to be considered significant then the p-values for the next LOD scores in line have to be estimated or until an unsignificant LOD score is found. The p-values are estimated via simulation. 265 unrelated individuals are simulated by drawing allele types independently with replacement from a gene pool with the same allele frequencies as the original dataset, excluding the foetuses. Then

their pairwise LOD scores are computed. The estimated p-value, $\hat{p}_{i,j}$, is the proportion of simulated LOD scores that are equal or higher than the original LOD score, $LOD_{i,j}$. $\hat{p}_{i,j}$ can be described as the estimated probability of attaining as extreme or more extreme LOD score than the original one, $LOD_{i,j}$, just by chance. The simulation procedure is replicated at least 60 times or until the standard errors for the $\hat{p}$-values, $sd(\hat{p}) = \sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{M}}$, are sufficiently small but M is the total number of simulated LOD scores. In this case sufficient means that the 95% confidence interval for the largest $\hat{p}$-value in the group of related pairs doesn't collide with the 95% confidence interval for the smallest $\hat{p}$-value in the group of unrelated pairs. The number of individuals simulated each time, 265, and the minimum number of replications of the procedure, 60, was chosen somewhat arbitrarily. It is the number of simulated LOD scores that matters, not the number of simulated individuals, and these numbers were convenient, computing time wise, to attain the needed number of simulated LOD scores.

In the third step a measure is taken to reduce the multiple comparison problem. In that step two methods for adjusting for multiple testing are applied and compared, the well known Bonferroni correction and Benjamini's and Hochberg's (1995) FDR procedure. In the Bonferroni procedure the $\hat{p}$-values are multiplied with the number of pairwise comparisons. The mother-foetus pairs are included in that number since it doesn't change the result for the non mother-foetus pairs whether they are included or not and it is of interest where the Bonferroni procedure places the mother-foetus pairs. The null hypothesis of unrelatedness is rejected for all pairs for which the Bonferroni adjusted $\hat{p}$-value is less or equal than $\alpha = 0.05$. The FDR procedure arranges the corresponding estimated p-values for each LOD score in an increasing order $\hat{p}_{(1)} \leq \hat{p}_{(2)} \leq .... \leq \hat{p}_{(n)}$. q is the target false discovery rate, n is the number of pairwise comparisons, and R is defined to be the largest value of r for which:

$$\hat{p}_{(r)} \leq \frac{r}{n} \cdot q = Q_r$$

The first R pairs in this sequence are classified as relatives and the remaining pairs are declared unrelated. The test procedure used in this analysis is based on having a large dataset of individuals of which little is known except for their DNA information. Therefore the mother-foetus pairs should not be included in the ranking of $\hat{p}$-values. However, it is of interest to see where the FDR procedure places the mother-foetus pairs and for that reason the FDR procedure is done twice, first including the mother-foetus pairs in the ranking and then without them. The false discovery rate is fixed at the same point as the significance level in the Bonferroni procedure, $q = \alpha = 0.05$. Having q and α of the same size simplifies the comparison of the two procedures but 0.05 is chosen to mirror the commonly chosen significance level in the single comparison case. Prior the result, the FDR procedure seems more suitable than the Bonferroni correction, see chapter 4.2.2, due to high number of LOD scores. The application of these two procedures, Bonferroni and FDR, should shed light on that.

All computations and simulations in the analysis were done by using the open source program R (R Development Core Team, 2011), but the codes can be found in appendix D

5.2 Data

The present study utilizes data from the Icelandic individual-based DNA registry of fin whales which compromises 267 genetic profiles collected between and during the years 2009 and 2010 and has been obtained for fifteen microsatellite loci, (EV001, EV037, GATA028, GATA053, GATA098, GATA417, GT011, GT023, GT195, GT211, GT271, GT310, GT575, TAA023 and GGAA520), the control region of mtDNA and a sex-marker. The age and age of maturity of the individuals within the database have been estimated by reading their plugs.

A total of 23 females, of the 267 indiviuals samples genotyped, carried a foetus for which a genetic sample was also obtained. Whales collected in 2009 were given a name starting with F09 and whales collected in 2010 were given a name starting with F10. The foetuses were given the same name as their mothers with the letter F applied to it at the end. Of those 290 genetic profiles, 265 are used in the present study. Information is missing at some loci for 24 individuals and one foetus so they were omitted from the analysis. One of the omitted individuals was a female that carried a foetus so there are 21 remaining mother-foetus pairs in the sample but 22 foetuses.

5.3 Population Allele Frequencies

The population allele frequencies were estimated directly from the sample, excluding the 22 foetuses since they can not be considered as part of the population. As was noted in chapter 3.2, Hardy-Weinberg and linkage equilibrium are assumed in this analysis and the fin whales within the sample are considered to belong to the same population. The estimation of the allele frequencies was done by dividing the number of times a certain allelic type was observed at a locus by the total number of alleles at that locus: $2 \cdot 243 = 486$. The computations were done in *R* (R Development Core Team, 2011) but the code can be found in appendix D.2. Tables with the estimated population frequencies values are in appendix E.

5.4 Application

5.4.1 Half-Siblings

Skaug et al. (2010) considered 'half-siblings' to be a reasonable choice for a general test to detect all types of close 1st- and 2nd order relationships. They pointed out that detection of parent-offspring dyads is highly sensitive for clerical errors but one typing error results in an infinitely negative LOD score. In the absence of an estimation of the error rate, LOD scores, based on 2nd-order dyads, were recommended for detecting both 1st- and 2nd-order relationships since they are more robust to typing errors. In the present study no typing error estimate has been attained and therefore the half-sibling LOD score is a good starting point to detect relatives within the dataset.

The half-sibling LOD score tests the hypothesis of half-siblings against the null hypothesis of unrelatedness:
H_0 : Individual i and individual j are unrelated

H_1 : Individual i and individual j are half-siblings

LOD Scores

A total of 34 980 pairwise LOD scores for the 265 individuals within the dataset were computed by using the formula:

$$\begin{aligned} LOD_{h.sib} =& log(\frac{P(D_i, D_j \mid half - siblings)}{P(D_i, D_j \mid not\ related)}) \\ =& log(\prod_{s=1}^{15}(\frac{1}{2} + \frac{1}{2} \cdot (\frac{I(a_{i,s}^{(1)} = a_{j,s}^{(1)}) + I(a_{i,s}^{(1)} = a_{j,s}^{(2)})}{4 \cdot p(a_{i,s}^{(1)})} \\ &+ \frac{I(a_{i,s}^{(2)} = a_{j,s}^{(1)}) + I(a_{i,s}^{(2)} = a_{j,s}^{(2)})}{4 \cdot p(a_{i,s}^{(2)})}))) \end{aligned}$$

Of those 34 980 scores, 3 731 are larger than zero, eleven are larger than 4 but the largest one is approximately 6.35.

The p-values were estimated by comparing LOD scores of simulated unrelated individuals with the original LOD scores. In this case it was sufficient to replicate the simulation procedure 60 times, resulting in 2 098 800 pair wise LOD scores of simulated unrelated individuals. The 1000 lowest $\hat{p}$-values, for the 1000 highest LOD scores, all look good in the single comparison case but the largest $\hat{p}$-value in that group is 0.031. If multiple testing was not taken into account one might just think that all of these pairs were dyads of relatives. Adjustment for multiple testing gives another result.

Table 5.1 shows the estimated p-values of the 50 highest half-siblings LOD scores, of which there are 19 mother-foetus pairs. The table also shows the standard errors of the estimated p-values. 95% confidence intervals were computed by using the standard normal approach for $\hat{p}$-values for which $M \cdot \hat{p} = 2\ 098\ 800 \cdot \hat{p} \geq 5$. That is sufficient for this analysis since confidence intervals are only needed for the two $\hat{p}$-values on the margin of related pairs and unrelated pairs of individuals.

Two mother-foetus pairs don't make it to the top 50 highest half-sibling LOD score list. If their data is examined it becomes apparent that F09-070 and F09-070F have not one identical allele type at the 6th locus and F10-067 and F10-067F do not have one identical allele type at the 14th locus. There is no way of knowing if this is because of mutation or because of an typing error when the data was sampled. Due to the high sensitivity of the parent-offspring LOD score it is known by forehand that the test procedure will not conclude these two mother-foetus pairs to be a mother and her offspring.

Bonferroni

The Bonferroni correction is attained by multiplying the $\hat{p}$-values with the number of pairwise LOD-scores computed from the dataset. Table 5.2 contains the 50 highest half-siblings LOD

Table 5.1: 50 highest pairwise half-sibling LOD scores and their corresponding $\hat{p}$-values

Pairs	$LOD_{h.sib}$	$\hat{p}$	$sd(\hat{p})$	95% confidence interval
F09-002 and F09-002F	6.34	0.00	0.00	$M \cdot \hat{p} < 5$
F10-073 and F10-073F	6.27	0.00	0.00	$M \cdot \hat{p} < 5$
F10-020 and F10-026	5.07	0.00	0.00	$M \cdot \hat{p} < 5$
F10-035 and F10-035F	4.61	0.00	0.00	$M \cdot \hat{p} < 5$
F09-073 and F10-062	4.58	0.00	0.00	$M \cdot \hat{p} < 5$
F10-018 and F10-018F	4.55	0.00	0.00	$M \cdot \hat{p} < 5$
F10-122 and F10-122F	4.43	0.00	0.00	$M \cdot \hat{p} < 5$
F10-134 and F10-134F	4.38	0.00	0.00	$M \cdot \hat{p} < 5$
F09-081 and F10-030	4.34	0.00	0.00	$M \cdot \hat{p} < 5$
F10-104 and F10-104F	4.30	0.00	0.00	$M \cdot \hat{p} < 5$
F10-044 and F10-044F	4.27	0.00	0.00	$M \cdot \hat{p} < 5$
F09-091 and F09-091F	3.94	$4.76 \cdot 10^{-7}$	$4.76 \cdot 10^{-7}$	$m \cdot \hat{p} < 5$
F09-047 and F10-079	3.70	$2.38 \cdot 10^{-6}$	$1.07 \cdot 10^{-6}$	$(2.94 \cdot 10^{-7}; 4.47 \cdot 10^{-6})$
F09-040 and F10-020	3.70	$2.86 \cdot 10^{-6}$	$1.17 \cdot 10^{-6}$	$(5.71 \cdot 10^{-7}; 5.15 \cdot 10^{-6})$
F10-089 and F10-140	3.64	$4.29 \cdot 10^{-6}$	$1.43 \cdot 10^{-6}$	$(1.49 \cdot 10^{-6}; 7.09 \cdot 10^{-6})$
F10-087 and F10-087F	3.62	$4.76 \cdot 10^{-6}$	$1.51 \cdot 10^{-6}$	$(1.81 \cdot 10^{-6}; 7.72 \cdot 10^{-6})$
F10-116 and F10-116F	3.58	$4.76 \cdot 10^{-6}$	$1.51 \cdot 10^{-6}$	$(1.81 \cdot 10^{-6}; 7.72 \cdot 10^{-6})$
F09-075 and F10-123	3.48	$6.19 \cdot 10^{-6}$	$1.72 \cdot 10^{-6}$	$(2.83 \cdot 10^{-6}; 9.56 \cdot 10^{-6})$
F09-091F and F10-100	3.42	$7.15 \cdot 10^{-6}$	$1.85 \cdot 10^{-6}$	$(3.53 \cdot 10^{-6}; 1.08 \cdot 10^{-5})$
F10-085 and F10-085F	3.27	$1.10 \cdot 10^{-5}$	$2.29 \cdot 10^{-6}$	$(6.48 \cdot 10^{-6}; 1.54 \cdot 10^{-5})$
F10-090 and F10-090F	3.26	$1.10 \cdot 10^{-5}$	$2.29 \cdot 10^{-6}$	$(6.48 \cdot 10^{-6}; 1.54 \cdot 10^{-5})$
F10-106 and F10-106F	3.16	$1.57 \cdot 10^{-5}$	$2.74 \cdot 10^{-6}$	$(1.04 \cdot 10^{-5}; 2.11 \cdot 10^{-5})$
F10-052 and F10-052F	3.14	$1.62 \cdot 10^{-5}$	$2.78 \cdot 10^{-6}$	$(1.08 \cdot 10^{-5}; 2.16 \cdot 10^{-5})$
F10-084 and F10-084F	3.11	$1.86 \cdot 10^{-5}$	$2.98 \cdot 10^{-6}$	$(1.28 \cdot 10^{-5}; 2.44 \cdot 10^{-5})$
F09-040 and F10-026	2.88	$4.81 \cdot 10^{-5}$	$4.79 \cdot 10^{-6}$	$(3.87 \cdot 10^{-5}; 5.75 \cdot 10^{-5})$
F09-125 and F10-119	2.84	$5.62 \cdot 10^{-5}$	$5.18 \cdot 10^{-6}$	$(4.61 \cdot 10^{-5}; 6.64 \cdot 10^{-5})$
F09-105 and F10-004	2.72	$8.39 \cdot 10^{-5}$	$6.32 \cdot 10^{-6}$	$(7.15 \cdot 10^{-5}; 9.62 \cdot 10^{-5})$
F09-095 and F09-107	2.72	$8.39 \cdot 10^{-5}$	$6.32 \cdot 10^{-6}$	$(7.15 \cdot 10^{-5}; 9.62 \cdot 10^{-5})$
F10-069 and F10-122F	2.67	$9.62 \cdot 10^{-5}$	$6.77 \cdot 10^{-6}$	$(8.30 \cdot 10^{-5}; 1.10 \cdot 10^{-4})$
F10-037 and F10-037F	2.62	$1.13 \cdot 10^{-4}$	$7.35 \cdot 10^{-6}$	$(9.90 \cdot 10^{-5}; 1.28 \cdot 10^{-4})$
F10-059 and F10-147	2.55	$1.48 \cdot 10^{-4}$	$8.39 \cdot 10^{-6}$	$(1.31 \cdot 10^{-4}; 1.64 \cdot 10^{-4})$
F09-065 and F10-004	2.50	$1.81 \cdot 10^{-4}$	$9.27 \cdot 10^{-6}$	$(1.62 \cdot 10^{-4}; 1.99 \cdot 10^{-4})$
F09-008 and F09-094	2.50	$1.82 \cdot 10^{-4}$	$9.32 \cdot 10^{-6}$	$(1.64 \cdot 10^{-4}; 2.01 \cdot 10^{-4})$
F10-026 and F10-099	2.48	$1.90 \cdot 10^{-4}$	$9.50 \cdot 10^{-6}$	$(1.71 \cdot 10^{-4}; 2.08 \cdot 10^{-4})$
F10-017 and F10-043	2.46	$2.05 \cdot 10^{-4}$	$9.88 \cdot 10^{-6}$	$(1.86 \cdot 10^{-4}; 2.24 \cdot 10^{-4})$
F10-073 and F10-097	2.41	$2.46 \cdot 10^{-4}$	$1.08 \cdot 10^{-5}$	$(2.25 \cdot 10^{-4}; 2.67 \cdot 10^{-4})$
F10-111 and F10-123	2.41	$2.48 \cdot 10^{-4}$	$1.09 \cdot 10^{-5}$	$(2.26 \cdot 10^{-4}; 2.69 \cdot 10^{-4})$
F09-040 and F10-135	2.38	$2.73 \cdot 10^{-4}$	$1.14 \cdot 10^{-5}$	$(2.50 \cdot 10^{-4}; 2.95 \cdot 10^{-4})$
F09-035 and F10-006	2.34	$3.08 \cdot 10^{-4}$	$1.21 \cdot 10^{-5}$	$(2.85 \cdot 10^{-4}; 3.32 \cdot 10^{-4})$
F09-007 and F10-042	2.31	$3.36 \cdot 10^{-4}$	$1.27 \cdot 10^{-5}$	$(3.12 \cdot 10^{-4}; 3.61 \cdot 10^{-4})$
F09-054 and F10-067	2.29	$3.66 \cdot 10^{-4}$	$1.32 \cdot 10^{-5}$	$(3.41 \cdot 10^{-4}; 3.92 \cdot 10^{-4})$
F10-022 and F10-022F	2.23	$4.39 \cdot 10^{-4}$	$1.45 \cdot 10^{-5}$	$(4.11 \cdot 10^{-4}; 4.68 \cdot 10^{-4})$
F09-044 and F10-062	2.23	$4.42 \cdot 10^{-4}$	$1.45 \cdot 10^{-5}$	$(4.14 \cdot 10^{-4}; 4.71 \cdot 10^{-4})$
F10-026 and F10-113	2.21	$4.79 \cdot 10^{-4}$	$1.51 \cdot 10^{-5}$	$(4.50 \cdot 10^{-4}; 5.09 \cdot 10^{-4})$
F10-111 and F10-111F	2.17	$5.37 \cdot 10^{-4}$	$1.60 \cdot 10^{-5}$	$(5.06 \cdot 10^{-4}; 5.69 \cdot 10^{-4})$
F09-081 and F10-106	2.17	$5.37 \cdot 10^{-4}$	$1.60 \cdot 10^{-5}$	$(5.06 \cdot 10^{-4}; 5.69 \cdot 10^{-4})$
F09-100 and F10-135	2.15	$5.77 \cdot 10^{-4}$	$1.66 \cdot 10^{-5}$	$(5.44 \cdot 10^{-4}; 6.09 \cdot 10^{-4})$
F09-021 and F10-086	2.10	$6.65 \cdot 10^{-4}$	$1.78 \cdot 10^{-5}$	$(6.30 \cdot 10^{-4}; 7.00 \cdot 10^{-4})$
F10-060 and F10-125	2.08	$7.22 \cdot 10^{-4}$	$1.85 \cdot 10^{-5}$	$(6.86 \cdot 10^{-4}; 7.58 \cdot 10^{-4})$
F09-116 and F10-111F	2.06	$7.60 \cdot 10^{-4}$	$1.90 \cdot 10^{-5}$	$(7.23 \cdot 10^{-4}; 7.97 \cdot 10^{-4})$

scores and their Bonferroni corrected $\hat{p}$-values. The Bonferroni procedure is very strict in the case of large number of pairwise comparisons. By using the Bonferroni adjustment and putting $\alpha = 0.05$, only twelve pairs of 34 980 are classified as relatives. Nine of them are mother-foetus pairs. Twelve mother-foetus pairs are concluded unrelated which demonstrates clearly how strict the Bonferroni procedure is.

FDR

The FDR procedure is based on the arrangement of the estimated p-values for each LOD score in an increasing order $\hat{p}_{(1)} \leq \hat{p}_{(2)} \leq .... \leq \hat{p}_{(n)}$. Table 5.3 contains information about the FDR procedure for the half-siblings hypothesis including the mother-foetus pairs in the ranking of $\hat{p}$-values. At $q = 0.05$, 24 of the pairs are concluded related and, as table 5.1 shows, the 95% confidence intervals of $\hat{p}_{(24)}$ and $\hat{p}_{(25)}$ don't collide. Sixteen of the related pairs are mother-foetus pairs and eight are non mother-foetus pairs. Five of the mother-foetus pairs within the dataset are concluded unrelated here while twelve of the mother-foetus pairs are concluded unrelated when the Bonferroni correction is used. That means that the FDR procedure detects at least seven more pairs of true relatives than the Bonferroni procedure. Not including the mother-foetus pairs in the ranking of the $\hat{p}$-values gives the same result for the non mother-foetus pairs. The eight pairs with the highest half-sibling LOD scores of the 34 959 non mother-foetus pairs are concluded related but the 95% confidence intervals of $\hat{p}_{(8)}$, (estimated p-value for the half-sibling LOD score of F09-091F and F10-100), and $\hat{p}_{(9)}$, (the estimated p-value for the half-sibling LOD score of F09-040 and F10-026), do not collide.

5.4.2 Parent-Offspring

The parent-offspring LOD score compares the hypothesis of parent-offspring relations against the null hypothesis of unrelatedness.

LOD Scores

A total of 34 980 parent-offspring LOD scores were calculated by using the formula:

$$\begin{aligned} LOD_p =& log(\frac{P(D_i, D_j \mid parent\ and\ offspring)}{P(D_i, D_j \mid not\ related)}) \\ =& log(\prod_{s=1}^{15} \frac{I(a_{i,s}^{(1)} = a_{j,s}^{(1)}) + I(a_{i,s}^{(1)} = a_{j,s}^{(2)})}{4 \cdot p(a_{i,s}^{(1)})} \\ &+ \frac{I(a_{i,s}^{(2)} = a_{j,s}^{(1)}) + I(a_{i,s}^{(2)} = a_{j,s}^{(2)})}{4 \cdot p(a_{i,s}^{(2)})}) \end{aligned}$$

Of these 34 980 scores, only 28 are not infinitely negative, there of are 19 mother-foetus pairs. As was mentioned in the half-siblings section, two of the mother-foetus pairs have an infinitely negative parent-offspring LOD scores.

The p-values were estimated by comparing LOD scores of simulated unrelated individuals with the original LOD scores. The simulation procedure was replicated 60 times, resulting

Table 5.2: 50 highest pairwise half-sibling LOD scores and their corresponding Bonferroni corrected $\hat{p}$-values

Pairs	$LOD_{h.sib}$	Bonferroni corrected $\hat{p}$	Decision at $\alpha = 0.05$
F09-002 and F09-002F	6.34	0.00	Related
F10-073 and F10-073F	6.27	0.00	Related
F10-020 and F10-026	5.07	0.00	Related
F10-035 and F10-035F	4.61	0.00	Related
F09-073 and F10-062	4.58	0.00	Related
F10-018 and F10-018F	4.55	0.00	Related
F10-122 and F10-122F	4.43	0.00	Related
F10-134 and F10-134F	4.38	0.00	Related
F09-081 and F10-030	4.34	0.00	Related
F10-104 and F10-104F	4.30	0.00	Related
F10-044 and F10-044F	4.27	0.00	Related
F09-091 and F09-091F	3.94	0.02	Related
F09-047 and F10-079	3.70	0.08	Unrelated
F09-040 and F10-020	3.70	0.10	Unrelated
F10-089 and F10-140	3.64	0.15	Unrelated
F10-087 and F10-087F	3.62	0.17	Unrelated
F10-116 and F10-116F	3.58	0.17	Unrelated
F09-075 and F10-123	3.48	0.22	Unrelated
F09-091F and F10-100	3.42	0.25	Unrelated
F10-085 and F10-085F	3.27	0.38	Unrelated
F10-090 andF10-090F	3.26	0.38	Unrelated
F10-106 and F10-106F	3.16	0.55	Unrelated
F10-052 and F10-052F	3.14	0.57	Unrelated
F10-084 and F10-084F	3.11	0.65	Unrelated
F09-040 and F10-026	2.88	1.68	Unrelated
F09-125 and F10-119	2.84	1.97	Unrelated
F09-105 and F10-004	2.72	2.93	Unrelated
F09-095 and F09-107	2.72	2.93	Unrelated
F10-069 and F10-122F	2.67	3.37	Unrelated
F10-037 and F10-37F	2.62	3.97	Unrelated
F10-059 and F10-147	2.55	5.17	Unrelated
F09-065 and F10-004	2.50	6.32	Unrelated
F09-008 and F09-094	2.50	6.38	Unrelated
F10-026 and F10-099	2.48	6.63	Unrelated
F10-017 and F10-043	2.46	7.17	Unrelated
F10-073 and F10-097	2.41	8.60	Unrelated
F10-111 and F10-123	2.41	8.67	Unrelated
F09-040 and F10-135	2.38	9.53	Unrelated
F09-035 and F10-006	2.34	10.78	Unrelated
F09-007 and F10-042	2.31	11.77	Unrelated
F09-054 and F10-067	2.29	12.82	Unrelated
F10-022 and F10-022F	2.23	15.37	Unrelated
F09-044 and F10-062	2.23	15.47	Unrelated
F10-026 and F10-113	2.21	16.77	Unrelated
F10-111 and F10-111F	2.17	18.80	Unrelated
F09-081 and F10-106	2.17	18.80	Unrelated
F09-100 and F10-135	2.15	20.17	Unrelated
F09-021 and F10-086	2.10	23.27	Unrelated
F10-060 and F10-125	2.08	25.25	Unrelated
F09-116 and F10-111F	2.06	26.58	Unrelated

Table 5.3: 50 highest pairwise half-sibling LOD scores and their corresponding Q_r-values

Pairs	$LOD_{h.sib}$	$\hat{p}$	r	$Q_r = (r/n) \cdot q$	Decision at $q = 0.05$
F09-002 and F09-002F	6.34	0.00	1	$1.43 \cdot 10^{-6}$	Related
F10-073 and F10-073F	6.27	0.00	2	$2.86 \cdot 10^{-6}$	Related
F10-020 and F10-026	5.07	0.00	3	$4.29 \cdot 10^{-6}$	Related
F10-035 and F10-035F	4.61	0.00	4	$5.72 \cdot 10^{-6}$	Related
F09-073 and F10-062	4.58	0.00	5	$7.15 \cdot 10^{-6}$	Related
F10-018 and F10-018F	4.55	0.00	6	$8.58 \cdot 10^{-6}$	Related
F10-122 and F10-122F	4.43	0.00	7	$1.00 \cdot 10^{-5}$	Related
F10-134 and F10-134F	4.38	0.00	8	$1.14 \cdot 10^{-5}$	Related
F09-081 and F10-030	4.34	0.00	9	$1.29 \cdot 10^{-5}$	Related
F10-104 and F10-104F	4.30	0.00	10	$1.43 \cdot 10^{-5}$	Related
F10-044 and F10-044F	4.27	0.00	11	$1.57 \cdot 10^{-5}$	Related
F09-091 and F09-091F	3.94	$4.76 \cdot 10^{-7}$	12	$1.72 \cdot 10^{-5}$	Related
F09-047 and F10-079	3.70	$2.38 \cdot 10^{-6}$	13	$1.86 \cdot 10^{-5}$	Related
F09-040 and F10-020	3.70	$2.86 \cdot 10^{-6}$	14	$2.00 \cdot 10^{-5}$	Related
F10-089 and F10-140	3.64	$4.29 \cdot 10^{-6}$	15	$2.14 \cdot 10^{-5}$	Related
F10-087 and F10-087F	3.62	$4.76 \cdot 10^{-6}$	16	$2.29 \cdot 10^{-5}$	Related
F10-116 and F10-116F	3.58	$4.76 \cdot 10^{-6}$	17	$2.43 \cdot 10^{-5}$	Related
F09-075 and F10-123	3.48	$6.19 \cdot 10^{-6}$	18	$2.57 \cdot 10^{-5}$	Related
F09-091F and F10-100	3.42	$7.15 \cdot 10^{-6}$	19	$2.72 \cdot 10^{-5}$	Related
F10-085 and F10-085F	3.27	$1.10 \cdot 10^{-5}$	20	$2.86 \cdot 10^{-5}$	Related
F10-090 andF10-090F	3.26	$1.10 \cdot 10^{-5}$	21	$3.00 \cdot 10^{-5}$	Related
F10-106 and F10-106F	3.16	$1.57 \cdot 10^{-5}$	22	$3.14 \cdot 10^{-5}$	Related
F10-052 and F10-052F	3.14	$1.62 \cdot 10^{-5}$	23	$3.29 \cdot 10^{-5}$	Related
F10-084 and F10-084F	3.11	$1.86 \cdot 10^{-5}$	24	$3.43 \cdot 10^{-5}$	Related
F09-040 and F10-026	2.88	$4.81 \cdot 10^{-5}$	25	$3.57 \cdot 10^{-5}$	Unrelated
F09-125 and F10-119	2.84	$5.62 \cdot 10^{-5}$	26	$3.72 \cdot 10^{-5}$	Unrelated
F09-105 and F10-004	2.72	$8.39 \cdot 10^{-5}$	27	$3.86 \cdot 10^{-5}$	Unrelated
F09-095 and F09-107	2.72	$8.39 \cdot 10^{-5}$	28	$4.00 \cdot 10^{-5}$	Unrelated
F10-069 and F10-122F	2.67	$9.62 \cdot 10^{-5}$	29	$4.15 \cdot 10^{-5}$	Unrelated
F10-037 and F10-37F	2.62	$1.13 \cdot 10^{-4}$	30	$4.29 \cdot 10^{-5}$	Unrelated
F10-059 and F10-147	2.55	$1.48 \cdot 10^{-4}$	31	$4.43 \cdot 10^{-5}$	Unrelated
F09-065 and F10-004	2.50	$1.81 \cdot 10^{-4}$	32	$4.57 \cdot 10^{-5}$	Unrelated
F09-008 and F09-094	2.50	$1.82 \cdot 10^{-4}$	33	$4.72 \cdot 10^{-5}$	Unrelated
F10-026 and F10-099	2.48	$1.90 \cdot 10^{-4}$	34	$4.86 \cdot 10^{-5}$	Unrelated
F10-017 and F10-043	2.46	$2.05 \cdot 10^{-4}$	35	$5.00 \cdot 10^{-5}$	Unrelated
F10-073 and F10-097	2.41	$2.46 \cdot 10^{-4}$	36	$5.15 \cdot 10^{-5}$	Unrelated
F10-111 and F10-123	2.41	$2.48 \cdot 10^{-4}$	37	$5.29 \cdot 10^{-5}$	Unrelated
F09-040 and F10-135	2.38	$2.73 \cdot 10^{-4}$	38	$5.43 \cdot 10^{-5}$	Unrelated
F09-035 and F10-006	2.34	$3.08 \cdot 10^{-4}$	39	$5.57 \cdot 10^{-5}$	Unrelated
F09-007 and F10-042	2.31	$3.36 \cdot 10^{-4}$	40	$5.72 \cdot 10^{-5}$	Unrelated
F09-054 and F10-067	2.29	$3.66 \cdot 10^{-4}$	41	$5.86 \cdot 10^{-5}$	Unrelated
F10-022 and F10-022F	2.23	$4.39 \cdot 10^{-4}$	42	$6.00 \cdot 10^{-5}$	Unrelated
F09-044 and F10-062	2.23	$4.42 \cdot 10^{-4}$	43	$6.15 \cdot 10^{-5}$	Unrelated
F10-026 and F10-113	2.21	$4.79 \cdot 10^{-4}$	44	$6.29 \cdot 10^{-5}$	Unrelated
F10-111 and F10-111F	2.17	$5.37 \cdot 10^{-4}$	45	$6.43 \cdot 10^{-5}$	Unrelated
F09-081 and F10-106	2.17	$5.37 \cdot 10^{-4}$	46	$6.58 \cdot 10^{-5}$	Unrelated
F09-100 and F10-135	2.15	$5.77 \cdot 10^{-4}$	47	$6.72 \cdot 10^{-5}$	Unrelated
F09-021 and F10-086	2.10	$6.65 \cdot 10^{-4}$	48	$6.86 \cdot 10^{-5}$	Unrelated
F10-060 and F10-125	2.08	$7.22 \cdot 10^{-4}$	49	$7.00 \cdot 10^{-5}$	Unrelated
F09-116 and F10-111F	2.06	$7.60 \cdot 10^{-4}$	50	$7.15 \cdot 10^{-5}$	Unrelated

in 2 098 800 pairwise simulated LOD scores. As would be expected the $\hat{p}$-values for the infinitely negative LOD scores were equal to 1. Table 5.4 contains information on the 28 pairs that have a finite parent-offspring LOD score. The standard errors of the estimated p-values have been computed and the asymptotic 95% confidence intervals of the $\hat{p}$-values for which $\hat{p} \cdot 2\ 098\ 800 \geq 5$.

The $\hat{p}$-values for the parent-offspring LOD scores are very small and if they were evaluated as in the single comparison case then the conclusion would be that all the 28 pairs were a parent and his/her offspring. Adjustment for multiple testing results in fewer rejections of the null hypothesis of unrelatedness.

Bonferroni

The $\hat{p}$-values are Bonferroni corrected by multiplying them with the number of pairwise LOD-scores computed from the dataset. Table 5.5 contains the Bonferroni adjusted $\hat{p}$-values for the 28 highest parent-offspring LOD scores and the test decisions based on a significance level of $\alpha = 0.05$. By using the Bonferroni correction 21 of the 28 pairs with a positive LOD score are classified as a parent and his/her offspring. Sixteen of those 21 pairs are mother-foetus pairs.

FDR

The FDR procedure is based on arranging the corresponding estimated p-values for each LOD score in increasing order $\hat{p}_{(1)} \leq \hat{p}_{(2)} \leq .... \leq \hat{p}_{(n)}$. Unlike the half-sibling test, in this case the FDR result for the non mother-foetus pairs depends on whether the mother-foetus pairs are included in the ranking of the estimated p-values or not. Table 5.6 contains information about the 28 highest parent-offspring LOD scores, their estimated p-values, corresponding Q_r-values and the FDR decision which is based on the comparison of $\hat{p}_{(r)}$ and Q_r. Table 5.7 contains the same information for the 9 highest parent-offspring LOD scores when the mother-foetus pairs are not included in the ranking.

If mother-foetus pairs are included in the ranking, all 28 pairs with LOD scores that are not infinitely negative are considered to consist of parents and their offspring at $q = 0.05$. The FDR procedure would therefore correctly conclude all the 19 mother-foetus pairs with a finite LOD score as a parent-offspring pair if nothing was known about them except their DNA information. This demonstrates how much stricter the Bonferroni procedure is but three of the mother-foetus pairs with a finite LOD_p score were concluded unrelated at $\alpha = 0.05$ when the estimated p-values were adjusted with the Bonferroni method. By using the FDR procedure it becomes less likely that related individuals are wrongly concluded unrelated.

When the mother-foetus pairs are not included in the ranking, then five pairs of 34 959 are classified as a parent and his/her offspring. In the ranking without mother-foetus pairs, $\hat{p}_{(5)}$ is the estimated p-value for the parent-offspring LOD score of F09-075 and F10-123 and $\hat{p}_{(6)}$ is the estimated p-value for the parent-offspring LOD score of F09-125 and F10-119. Since $\hat{p}_{(5)} \cdot 2\ 098\ 800 < 5$, which means that its confidence interval should not be computed with the standard normal approach, the 95% exact binomial confidence intervals for $\hat{p}_{(5)}$ and $\hat{p}_{(6)}$ are computed by using the package binom (Dorai-Raj, 2009) in R, see appendix

Table 5.4: 28 highest pairwise parent-offspring LOD scores and their corresponding $\hat{p}$-values.

Pairs	LOD_p	$\hat{p}$	$sd(\hat{p})$	95% confidence interval
F09-002 and F09-002F	8.97	0.00	0.00	$M \cdot \hat{p} < 5$
F10-073 and F10-073F	8.46	0.00	0.00	$M \cdot \hat{p} < 5$
F10-020 and F10-026	7.40	0.00	0.00	$M \cdot \hat{p} < 5$
F10-035 and F10-035F	6.91	0.00	0.00	$M \cdot \hat{p} < 5$
F10-018 and F10-018F	6.85	0.00	0.00	$M \cdot \hat{p} < 5$
F09-081 and F10-030	6.68	0.00	0.00	$M \cdot \hat{p} < 5$
F10-104 and F10-104F	6.48	0.00	0.00	$M \cdot \hat{p} < 5$
F10-134 and F10-134F	6.48	0.00	0.00	$M \cdot \hat{p} < 5$
F10-122 and F10-122F	6.38	0.00	0.00	$M \cdot \hat{p} < 5$
F10-044 and F10-044F	6.33	0.00	0.00	$M \cdot \hat{p} < 5$
F09-091 and F09-091F	5.95	0.00	0.00	$M \cdot \hat{p} < 5$
F10-089 and F10-140	5.47	$4.76 \cdot 10^{-7}$	$4.76 \cdot 10^{-7}$	$M \cdot \hat{p} < 5$
F10-087 and F10-087F	5.46	$4.76 \cdot 10^{-7}$	$4.76 \cdot 10^{-7}$	$M \cdot \hat{p} < 5$
F09-91F and F10-100	5.43	$4.76 \cdot 10^{-7}$	$4.76 \cdot 10^{-7}$	$M \cdot \hat{p} < 5$
F09-075 and F10-123	5.38	$4.76 \cdot 10^{-7}$	$4.76 \cdot 10^{-7}$	$M \cdot \hat{p} < 5$
F10-116 and F10-116F	5.26	$4.76 \cdot 10^{-7}$	$4.76 \cdot 10^{-7}$	$M \cdot \hat{p} < 5$
F10-090 andF10-090F	5.07	$4.76 \cdot 10^{-7}$	$4.76 \cdot 10^{-7}$	$M \cdot \hat{p} < 5$
F10-106 and F10-106F	4.88	$4.76 \cdot 10^{-7}$	$4.76 \cdot 10^{-7}$	$M \cdot \hat{p} < 5$
F10-085 and F10-085F	4.87	$4.76 \cdot 10^{-7}$	$4.76 \cdot 10^{-7}$	$M \cdot \hat{p} < 5$
F10-052 and F10-052F	4.84	$9.53 \cdot 10^{-7}$	$6.74 \cdot 10^{-7}$	$M \cdot \hat{p} < 5$
F10-084 and F10-084F	4.79	$1.43 \cdot 10^{-6}$	$8.25 \cdot 10^{-7}$	$M \cdot \hat{p} < 5$
F10-037 and F10-37F	4.19	$7.15 \cdot 10^{-6}$	$1.85 \cdot 10^{-6}$	$(3.53 \cdot 10^{-6}; 1.08 \cdot 10^{-5})$
F09-125 and F10-119	4.05	$9.05 \cdot 10^{-6}$	$2.08 \cdot 10^{-6}$	$(4.98 \cdot 10^{-6}; 1.31 \cdot 10^{-5})$
F10-022 and F10-022F	3.67	$1.57 \cdot 10^{-5}$	$2.74 \cdot 10^{-6}$	$(1.04 \cdot 10^{-5}; 2.11 \cdot 10^{-5})$
F10-111 and F10-111F	3.42	$1.95 \cdot 10^{-5}$	$3.05 \cdot 10^{-6}$	$(1.36 \cdot 10^{-5}; 2.55 \cdot 10^{-5})$
F09-021 and F10-086	3.15	$2.38 \cdot 10^{-5}$	$3.37 \cdot 10^{-6}$	$(1.72 \cdot 10^{-5}; 3.04 \cdot 10^{-5})$
F09-068 and F10-146	3.10	$2.53 \cdot 10^{-5}$	$3.47 \cdot 10^{-6}$	$(1.85 \cdot 10^{-5}; 3.21 \cdot 10^{-5})$
F10-060 og F10-148	3.10	$2.53 \cdot 10^{-5}$	$3.47 \cdot 10^{-6}$	$(1.85 \cdot 10^{-5}; 3.21 \cdot 10^{-5})$

Table 5.5: 28 highest pairwise parent-offspring LOD scores and their corresponding Bonferroni corrected $\hat{p}$-values.

Names	LOD_p	Bonferroni corrected $\hat{p}$	Decision at $\alpha = 0.05$
F09-002 and F09-002F	8.97	0.00	Parent and offspring
F10-073 and F10-073F	8.46	0.00	Parent and offspring
F10-020 and F10-026	7.40	0.00	Parent and offspring
F10-035 and F10-035F	6.91	0.00	Parent and offspring
F10-018 and F10-018F	6.85	0.00	Parent and offspring
F09-081 and F10-030	6.68	0.00	Parent and offspring
F10-104 and F10-104F	6.48	0.00	Parent and offspring
F10-134 and F10-134F	6.48	0.00	Parent and offspring
F10-122 and F10-122F	6.38	0.00	Parent and offspring
F10-044 and F10-044F	6.33	0.00	Parent and offspring
F09-091 and F09-091F	5.95	0.00	Parent and offspring
F10-089 and F10-140	5.47	0.02	Parent and offspring
F10-087 and F10-087F	5.46	0.02	Parent and offspring
F09-091F and F10-100	5.43	0.02	Parent and offspring
F09-075 and F10-123	5.38	0.02	Parent and offspring
F10-116 and F10-116F	5.26	0.02	Parent and offspring
F10-090 andF10-090F	5.07	0.02	Parent and offspring
F10-106 and F10-106F	4.88	0.02	Parent and offspring
F10-085 and F10-085F	4.87	0.02	Parent and offspring
F10-052 and F10-052F	4.84	0.03	Parent and offspring
F10-084 and F10-084F	4.79	0.05	Parent and offspring
F10-037 and F10-037F	4.19	0.25	Unrelated
F09-125 and F10-119	4.05	0.32	Unrelated
F10-022 and F10-022F	3.67	0.55	Unrelated
F10-111 and F10-111F	3.42	0.68	Unrelated
F09-021 and F10-086	3.15	0.83	Unrelated
F09-068 and F10-146	3.10	0.88	Unrelated
F10-060 og F10-148	3.10	0.88	Unrelated

Table 5.6: 28 highest pairwise parent-offspring LOD scores and their corresponding Q_r-values

Pairs	LOD_p	$\hat{p}$	r	$Q_r = (r/n) \cdot q$	Decision at $q = 0.05$
F09-002 and F09-002F	8.97	0.00	1	$1.43 \cdot 10^{-6}$	Parent-offspring
F10-073 and F10-073F	8.46	0.00	2	$2.86 \cdot 10^{-6}$	Parent-offspring
F10-020 and F10-026	7.40	0.00	3	$4.29 \cdot 10^{-6}$	Parent-offspring
F10-035 and F10-035F	6.91	0.00	4	$5.72 \cdot 10^{-6}$	Parent-offspring
F10-018 and F10-018F	6.85	0.00	5	$7.15 \cdot 10^{-6}$	Parent-offspring
F09-081 and F10-030	6.68	0.00	6	$8.58 \cdot 10^{-6}$	Parent-offspring
F10-104 and F10-104F	6.48	0.00	7	$1.00 \cdot 10^{-5}$	Parent-offspring
F10-134 and F10-134F	6.48	0.00	8	$1.14 \cdot 10^{-5}$	Parent-offspring
F10-122 and F10-122F	6.38	0.00	9	$1.29 \cdot 10^{-5}$	Parent-offspring
F10-044 and F10-044F	6.33	0.00	10	$1.43 \cdot 10^{-5}$	Parent-offspring
F09-091 and F09-091F	5.95	0.00	11	$1.57 \cdot 10^{-5}$	Parent-offspring
F10-089 and F10-140	5.47	$4.76 \cdot 10^{-7}$	12	$1.72 \cdot 10^{-5}$	Parent-offspring
F10-087 and F10-087F	5.46	$4.76 \cdot 10^{-7}$	13	$1.86 \cdot 10^{-5}$	Parent-offspring
F09-91F and F10-100	5.43	$4.76 \cdot 10^{-7}$	14	$2.00 \cdot 10^{-5}$	Parent-offspring
F09-075 and F10-123	5.38	$4.76 \cdot 10^{-7}$	15	$2.14 \cdot 10^{-5}$	Parent-offspring
F10-116 and F10-116F	5.26	$4.76 \cdot 10^{-7}$	16	$2.29 \cdot 10^{-5}$	Parent-offspring
F10-090 andF10-090F	5.07	$4.76 \cdot 10^{-7}$	17	$2.43 \cdot 10^{-5}$	Parent-offspring
F10-106 and F10-106F	4.88	$4.76 \cdot 10^{-7}$	18	$2.57 \cdot 10^{-5}$	Parent-offspring
F10-085 and F10-085F	4.87	$4.76 \cdot 10^{-7}$	19	$2.72 \cdot 10^{-5}$	Parent-offspring
F10-052 and F10-052F	4.84	$9.53 \cdot 10^{-7}$	20	$2.86 \cdot 10^{-5}$	Parent-offspring
F10-084 and F10-084F	4.79	$1.43 \cdot 10^{-6}$	21	$3.00 \cdot 10^{-5}$	Parent-offspring
F10-037 and F10-37F	4.19	$7.15 \cdot 10^{-6}$	22	$3.14 \cdot 10^{-5}$	Parent-offspring
F09-125 and F10-119	4.05	$9.05 \cdot 10^{-6}$	23	$3.29 \cdot 10^{-5}$	Parent-offspring
F10-022 and F10-022F	3.67	$1.57 \cdot 10^{-5}$	24	$3.43 \cdot 10^{-5}$	Parent-offspring
F10-111 and F10-111F	3.42	$1.95 \cdot 10^{-5}$	25	$3.57 \cdot 10^{-5}$	Parent-offspring
F09-021 and F10-086	3.15	$2.38 \cdot 10^{-5}$	26	$3.72 \cdot 10^{-5}$	Parent-offspring
F09-068 and F10-146	3.10	$2.53 \cdot 10^{-5}$	27	$3.86 \cdot 10^{-5}$	Parent-offspring
F10-060 og F10-148	3.10	$2.53 \cdot 10^{-5}$	28	$4.00 \cdot 10^{-5}$	Parent-offspring

Table 5.7: 9 highest pairwise parent-offspring LOD scores, mother-foetus pairs not included, and their corresponding Q_r-values

Pairs	LOD_p	$\hat{p}$	r	$Q_r = (r/n) \cdot q$	Decision at $q = 0.05$
F10-020 and F10-026	7.40	0.00	1	$1.43 \cdot 10^{-6}$	Parent-offspring
F09-081 and F10-030	6.68	0.00	2	$2.86 \cdot 10^{-6}$	Parent-offspring
F10-089 and F10-140	5.47	$4.76 \cdot 10^{-7}$	3	$4.29 \cdot 10^{-6}$	Parent-offspring
F09-091F and F10-100	5.43	$4.76 \cdot 10^{-7}$	4	$5.72 \cdot 10^{-6}$	Parent-offspring
F09-075 and F10-123	5.38	$4.76 \cdot 10^{-7}$	5	$7.15 \cdot 10^{-6}$	Parent-offspring
F09-125 and F10-119	4.05	$9.05 \cdot 10^{-6}$	6	$8.58 \cdot 10^{-6}$	Unrelated
F09-021 and F10-086	3.15	$2.38 \cdot 10^{-5}$	7	$1.00 \cdot 10^{-5}$	Unrelated
F98-068 and F10-146	3.10	$2.53 \cdot 10^{-5}$	8	$1.14 \cdot 10^{-5}$	Unrelated
F10-060 og F10-148	3.10	$2.53 \cdot 10^{-5}$	9	$1.29 \cdot 10^{-5}$	Unrelated

D.6. The 95% exact binomial confidence intervals for $\hat{p}_{(5)}$, $(1.21 \cdot 10^{-8}; 2.65 \cdot 10^{-6})$, and $\hat{p}_{(6)}$, $(5.45 \cdot 10^{-6}; 1.41 \cdot 10^{-5})$, do not collide.

One of those five classified parent-offspring pairs consists of a foetus, F09-091F, and a male fin whale, F10-100. In this specific case, auxiliary genetic data is available, the DNA-profile of the mother, F09-091. When the profile of F09-091 is taken into account there is still a match between F09-091F and F10-100. It looks like the father of foetus F09-091F has been found. This will be examined closer in the result section.

5.4.3 First Cousins

The first cousins LOD score compares the hypothesis of a first cousins relations against the null hypothesis of unrelatedness.

LOD scores

A total of 34 980 first cousins LOD scores were calculated by using the formula:

$$\begin{aligned} LOD_{h.sib} &= log(\frac{P(D_i, D_j \mid first\ cousins)}{P(D_i, D_j \mid not\ related)}) \\ &= log(\prod_{s=1}^{15}(\frac{3}{4} + \frac{1}{4} \cdot (\frac{I(a_{i,s}^{(1)} = a_{j,s}^{(1)}) + I(a_{i,s}^{(1)} = a_{j,s}^{(2)})}{4 \cdot p(a_{i,s}^{(1)})} \\ &\quad + \frac{I(a_{i,s}^{(2)} = a_{j,s}^{(1)}) + I(a_{i,s}^{(2)} = a_{j,s}^{(2)})}{4 \cdot p(a_{i,s}^{(2)})}))) \end{aligned}$$

Of those 34 980 LOD scores, 8 238 are larger than zero, 23 are larger than 2 but the largest one is approximately 4.48. The $\hat{p}$-values for each LOD score were attained by simulating 265 unrelated individuals, computing their pair wise first cousin LOD scores and comparing them with the original LOD scores. The simulation procedure had to be repeated 90 times, resulting in 3 148 200 simulated LOD scores, in order to attain sufficiently small confidence intervals for the estimated p-values. Table 5.8 contains the fifty highest first cousins LOD scores and their corresponding $\hat{p}$-values. Standard deviation was computed for each $\hat{p}$ as well as 95% asymptotic confidence intervals for all $\hat{p}$ that satisfy $M \cdot \hat{p} \geq 5$.

As with the half-sibling test procedure, the 1000 lowest $\hat{p}$-values of the 1000 highest first cousins LOD scores all look good in the single comparison case but the largest estimated p-value in that group is approximately 0.031. If the problem with multiple testing was not taken into account then at least 1000 pairs would be classified as first cousins at $\alpha = 0.05$. Adjustment for multiple testing reduces the number of rejections of the null hypothesis.

Bonferroni

The Bonferroni adjusted $\hat{p}$ values are simply attained by multiplying the original $\hat{p}$-values with the number of pairwise LOD scores, 34 980. The mother-foetus pairs are included in that number since it doesn't change the result for the non mother-foetus pairs whether they are included or not. Table 5.9 contains the 50 highest first cousins LOD scores and their

Table 5.8: 50 highest pairwise first cousins LOD scores and their corresponding $\hat{p}$-values

Pairs	LOD_{cous}	$\hat{p}$	$sd(\hat{p})$	95% confidence interval
F10-073 and F10-073F	4.48	0.00	0.00	$M \cdot \hat{p} < 5$
F09-002 and F09-002F	4.30	0.00	0.00	$M \cdot \hat{p} < 5$
F10-020 and F10-026	3.28	0.00	0.00	$M \cdot \hat{p} < 5$
F09-073 and F10-062	3.00	0.00	0.00	$M \cdot \hat{p} < 5$
F10-122 and F10-122F	2.90	$6.35 \cdot 10^{-7}$	$4.49 \cdot 10^{-7}$	$M \cdot \hat{p} < 5$
F10-035 and F10-035F	2.87	$9.53 \cdot 10^{-7}$	$5.50 \cdot 10^{-7}$	$M \cdot \hat{p} < 5$
F10-018 and F10-018F	2.86	$9.53 \cdot 10^{-7}$	$5.50 \cdot 10^{-7}$	$M \cdot \hat{p} < 5$
F10-134 and F10-134F	2.78	$1.27 \cdot 10^{-6}$	$6.35 \cdot 10^{-7}$	$M \cdot \hat{p} < 5$
F09-047 and F10-079	2.78	$1.27 \cdot 10^{-6}$	$6.35 \cdot 10^{-7}$	$M \cdot \hat{p} < 5$
F10-044 and F10-044F	2.70	$2.22 \cdot 10^{-6}$	$8.40 \cdot 10^{-7}$	$(5.76 \cdot 10^{-7}; 3.87 \cdot 10^{-6})$
F10-104 and F10-104F	2.68	$2.54 \cdot 10^{-6}$	$8.98 \cdot 10^{-7}$	$(7.80 \cdot 10^{-7}; 4.30 \cdot 10^{-6})$
F09-081 and F10-030	2.65	$2.86 \cdot 10^{-6}$	$9.53 \cdot 10^{-7}$	$(9.91 \cdot 10^{-7}; 4.73 \cdot 10^{-6})$
F09-040 and F10-020	2.59	$3.81 \cdot 10^{-6}$	$1.10 \cdot 10^{-6}$	$(1.66 \cdot 10^{-6}; 5.97 \cdot 10^{-6})$
F09-091 and F09-091F	2.45	$6.99 \cdot 10^{-6}$	$1.49 \cdot 10^{-6}$	$(4.07 \cdot 10^{-6}; 9.91 \cdot 10^{-6})$
F10-116 and F10-116F	2.31	$1.49 \cdot 10^{-5}$	$2.18 \cdot 10^{-6}$	$(1.07 \cdot 10^{-5}; 1.92 \cdot 10^{-5})$
F10-089 and F10-140	2.28	$1.91 \cdot 10^{-5}$	$2.46 \cdot 10^{-6}$	$(1.42 \cdot 10^{-5}; 2.39 \cdot 10^{-5})$
F10-087 and F10-087F	2.22	$2.22 \cdot 10^{-5}$	$2.66 \cdot 10^{-6}$	$(1.70 \cdot 10^{-5}; 2.74 \cdot 10^{-5})$
F09-105 and F10-004	2.14	$3.43 \cdot 10^{-5}$	$3.30 \cdot 10^{-6}$	$(2.78 \cdot 10^{-5}; 4.08 \cdot 10^{-5})$
F09-075 and F10-123	2.13	$3.49 \cdot 10^{-5}$	$3.33 \cdot 10^{-6}$	$(2.84 \cdot 10^{-5}; 4.15 \cdot 10^{-5})$
F10-069 and F10-122F	2.10	$4.16 \cdot 10^{-5}$	$3.64 \cdot 10^{-6}$	$(3.45 \cdot 10^{-5}; 4.87 \cdot 10^{-5})$
F09-040 and F10-026	2.10	$4.16 \cdot 10^{-5}$	$3.64 \cdot 10^{-6}$	$(3.45 \cdot 10^{-5}; 4.87 \cdot 10^{-5})$
F10-085 and F10-085F	2.05	$4.96 \cdot 10^{-5}$	$3.97 \cdot 10^{-6}$	$(4.18 \cdot 10^{-5}; 5.73 \cdot 10^{-5})$
F09-91F and F10-100	2.00	$6.35 \cdot 10^{-5}$	$4.49 \cdot 10^{-6}$	$(5.47 \cdot 10^{-5}; 7.23 \cdot 10^{-5})$
F10-090 andF10-090F	1.95	$8.29 \cdot 10^{-5}$	$5.13 \cdot 10^{-6}$	$(7.28 \cdot 10^{-5}; 9.30 \cdot 10^{-5})$
F10-052 and F10-052F	1.92	$9.18 \cdot 10^{-5}$	$5.40 \cdot 10^{-6}$	$(8.12 \cdot 10^{-5}; 1.02 \cdot 10^{-4})$
F09-095 and F09-107	1.91	$9.66 \cdot 10^{-5}$	$5.54 \cdot 10^{-6}$	$(8.57 \cdot 10^{-5}; 1.07 \cdot 10^{-4})$
F10-106 and F10-106F	1.91	$9.66 \cdot 10^{-5}$	$5.54 \cdot 10^{-6}$	$(8.57 \cdot 10^{-5}; 1.07 \cdot 10^{-4})$
F10-084 and F10-084F	1.85	$1.38 \cdot 10^{-4}$	$6.62 \cdot 10^{-6}$	$(1.25 \cdot 10^{-4}; 1.51 \cdot 10^{-4})$
F09-125 and F10-119	1.81	$1.75 \cdot 10^{-4}$	$7.46 \cdot 10^{-6}$	$(1.60 \cdot 10^{-4}; 1.90 \cdot 10^{-4})$
F10-026 and F10-099	1.75	$2.33 \cdot 10^{-4}$	$8.59 \cdot 10^{-6}$	$(2.16 \cdot 10^{-4}; 2.49 \cdot 10^{-4})$
F10-059 and F10-147	1.68	$3.32 \cdot 10^{-4}$	$1.03 \cdot 10^{-5}$	$(3.12 \cdot 10^{-4}; 3.52 \cdot 10^{-4})$
F09-065 and F10-004	1.67	$3.47 \cdot 10^{-4}$	$1.05 \cdot 10^{-5}$	$(3.27 \cdot 10^{-4}; 3.68 \cdot 10^{-4})$
F10-027 and F10-142	1.64	$3.83 \cdot 10^{-4}$	$1.10 \cdot 10^{-5}$	$(3.61 \cdot 10^{-4}; 4.05 \cdot 10^{-4})$
F09-035 and F10-006	1.62	$4.37 \cdot 10^{-4}$	$1.18 \cdot 10^{-5}$	$(4.14 \cdot 10^{-4}; 4.60 \cdot 10^{-4})$
F09-103 and F10-017	1.61	$4.43 \cdot 10^{-4}$	$1.19 \cdot 10^{-5}$	$(4.20 \cdot 10^{-4}; 4.66 \cdot 10^{-4})$
F10-073 and F10-097	1.60	$4.68 \cdot 10^{-4}$	$1.22 \cdot 10^{-5}$	$(4.44 \cdot 10^{-4}; 4.92 \cdot 10^{-4})$
F10-111 and F10-123	1.59	$4.85 \cdot 10^{-4}$	$1.24 \cdot 10^{-5}$	$(4.60 \cdot 10^{-4}; 5.09 \cdot 10^{-4})$
F09-044 and F10-062	1.58	$5.21 \cdot 10^{-4}$	$1.29 \cdot 10^{-5}$	$(4.96 \cdot 10^{-4}; 5.46 \cdot 10^{-4})$
F09-008 and F09-094	1.56	$5.56 \cdot 10^{-4}$	$1.33 \cdot 10^{-5}$	$(5.30 \cdot 10^{-4}; 5.82 \cdot 10^{-4})$
F10-017 and F10-043	1.55	$5.97 \cdot 10^{-4}$	$1.38 \cdot 10^{-5}$	$(5.70 \cdot 10^{-4}; 6.24 \cdot 10^{-4})$
F10-037 and F10-37F	1.53	$6.50 \cdot 10^{-4}$	$1.44 \cdot 10^{-5}$	$(6.22 \cdot 10^{-4}; 6.78 \cdot 10^{-4})$
F09-040 and F10-135	1.51	$7.05 \cdot 10^{-4}$	$1.50 \cdot 10^{-5}$	$(6.76 \cdot 10^{-4}; 7.35 \cdot 10^{-4})$
F09-069 and FF10-006	1.49	$7.84 \cdot 10^{-4}$	$1.58 \cdot 10^{-5}$	$(7.53 \cdot 10^{-4}; 8.15 \cdot 10^{-4})$
F09-054 and F10-067	1.48	$8.19 \cdot 10^{-4}$	$1.61 \cdot 10^{-5}$	$(7.87 \cdot 10^{-4}; 8.50 \cdot 10^{-4})$
F09-011 and F09-124	1.48	$8.19 \cdot 10^{-4}$	$1.61 \cdot 10^{-5}$	$(7.87 \cdot 10^{-4}; 8.50 \cdot 10^{-4})$
F10-034 and F10-115	1.48	$8.22 \cdot 10^{-4}$	$1.62 \cdot 10^{-5}$	$(7.91 \cdot 10^{-4}; 8.54 \cdot 10^{-4})$
F10-059 and F10-078	1.47	$8.73 \cdot 10^{-4}$	$1.66 \cdot 10^{-5}$	$(8.40 \cdot 10^{-4}; 9.05 \cdot 10^{-4})$
F09-050 and F09-100	1.45	$9.27 \cdot 10^{-4}$	$1.72 \cdot 10^{-5}$	$(8.93 \cdot 10^{-4}; 9.60 \cdot 10^{-4})$
F10-026 and F10-113	1.45	$9.34 \cdot 10^{-4}$	$1.72 \cdot 10^{-5}$	$(9.00 \cdot 10^{-4}; 9.68 \cdot 10^{-4})$
F09-038 and F09-050	1.44	$9.72 \cdot 10^{-4}$	$1.76 \cdot 10^{-5}$	$(9.38 \cdot 10^{-4}; 1.01 \cdot 10^{-3})$

Table 5.9: 50 highest pairwise first cousins LOD scores and their Bonferroni corrected $\hat{p}$-values

Names	LOD_{cous}	Bonferroni corrected $\hat{p}$	Decision at $\alpha = 0.05$
F10-073 and F10-073F	4.48	0.00	First cousins
F09-002 and F09-002F	4.30	0.00	First cousins
F10-020 and F10-026	3.28	0.00	First cousins
F09-073 and F10-062	3.00	0.00	First cousins
F10-122 and F10-122F	2.90	0.02	First cousins
F10-035 and F10-035F	2.87	0.03	First cousins
F10-018 and F10-018F	2.86	0.03	First cousins
F10-134 and F10-134F	2.78	0.04	First cousins
F09-047 and F10-079	2.78	0.04	First cousins
F10-044 and F10-044F	2.70	0.08	Unrelated
F10-104 and F10-104F	2.68	0.09	Unrelated
F09-081 and F10-030	2.65	0.10	Unrelated
F09-040 and F10-020	2.59	0.13	Unrelated
F09-091 and F09-091F	2.45	0.24	Unrelated
F10-116 and F10-116F	2.31	0.52	Unrelated
F10-089 and F10-140	2.28	0.67	Unrelated
F10-087 and F10-087F	2.22	0.78	Unrelated
F09-105 and F10-004	2.14	1.20	Unrelated
F09-075 and F10-123	2.13	1.22	Unrelated
F10-069 and F10-122F	2.10	1.46	Unrelated
F09-040 and F10-026	2.10	1.46	Unrelated
F10-085 and F10-085F	2.05	1.73	Unrelated
F09-91F and F10-100	2.00	2.22	Unrelated
F10-090 andF10-090F	1.95	2.90	Unrelated
F10-052 and F10-052F	1.92	3.21	Unrelated
F09-095 and F09-107	1.91	3.38	Unrelated
F10-106 and F10-106F	1.91	3.38	Unrelated
F10-084 and F10-084F	1.85	4.82	Unrelated
F09-125 and F10-119	1.81	6.12	Unrelated
F10-026 and F10-099	1.75	8.13	Unrelated
F10-059 and F10-147	1.68	11.61	Unrelated
F09-065 and F10-004	1.67	12.14	Unrelated
F10-027 and F10-142	1.64	13.40	Unrelated
F09-035 and F10-006	1.62	15.28	Unrelated
F09-103 and F10-017	1.61	15.50	Unrelated
F10-073 and F10-097	1.60	16.38	Unrelated
F10-111 and F10-123	1.59	16.96	Unrelated
F09-044 and F10-062	1.58	18.22	Unrelated
F09-008 and F09-094	1.56	19.46	Unrelated
F10-017 and F10-043	1.55	20.87	Unrelated
F10-037 and F10-37F	1.53	22.74	Unrelated
F09-040 and F10-135	1.51	24.68	Unrelated
F09-069 and FF10-006	1.49	27.43	Unrelated
F09-054 and F10-067	1.48	28.63	Unrelated
F09-011 and F09-124	1.48	28.63	Unrelated
F10-034 and F10-115	1.48	28.77	Unrelated
F10-059 and F10-078	1.47	30.52	Unrelated
F09-050 and F09-100	1.45	32.42	Unrelated
F10-026 and F10-113	1.45	32.67	Unrelated
F09-038 and F09-050	1.44	34.00	Unrelated

Table 5.10: 50 highest pairwise first cousins LOD scores and their corresponding Q_r values

Pairs	LOD_{cous}	$\hat{p}$	r	$Q_r = (r/n) \cdot q$	Decision at $q = 0.05$
F10-073 and F10-073F	4.48	0.00	1	$1.43 \cdot 10^{-6}$	First cousins
F09-002 and F09-002F	4.30	0.00	2	$2.86 \cdot 10^{-6}$	First cousins
F10-020 and F10-026	3.28	0.00	3	$4.29 \cdot 10^{-6}$	First cousins
F09-073 and F10-062	3.00	0.00	4	$5.72 \cdot 10^{-6}$	First cousins
F10-122 and F10-122F	2.90	$6.35 \cdot 10^{-7}$	5	$7.15 \cdot 10^{-6}$	First cousins
F10-035 and F10-035F	2.87	$9.53 \cdot 10^{-7}$	6	$8.58 \cdot 10^{-6}$	First cousins
F10-018 and F10-018F	2.86	$9.53 \cdot 10^{-7}$	7	$1.00 \cdot 10^{-5}$	First cousins
F10-134 and F10-134F	2.78	$1.27 \cdot 10^{-6}$	8	$1.14 \cdot 10^{-5}$	First cousins
F09-047 and F10-079	2.78	$1.27 \cdot 10^{-6}$	9	$1.29 \cdot 10^{-5}$	First cousins
F10-044 and F10-044F	2.70	$2.22 \cdot 10^{-6}$	10	$1.43 \cdot 10^{-5}$	First cousins
F10-104 and F10-104F	2.68	$2.54 \cdot 10^{-6}$	11	$1.57 \cdot 10^{-5}$	First cousins
F09-081 and F10-030	2.65	$2.86 \cdot 10^{-6}$	12	$1.72 \cdot 10^{-5}$	First cousins
F09-040 and F10-020	2.59	$3.81 \cdot 10^{-6}$	13	$1.86 \cdot 10^{-5}$	First cousins
F09-091 and F09-091F	2.45	$6.99 \cdot 10^{-6}$	14	$2.00 \cdot 10^{-5}$	First cousins
F10-116 and F10-116F	2.31	$1.49 \cdot 10^{-5}$	15	$2.14 \cdot 10^{-5}$	First cousins
F10-089 and F10-140	2.28	$1.91 \cdot 10^{-5}$	16	$2.29 \cdot 10^{-5}$	First cousins
F10-087 and F10-087F	2.22	$2.22 \cdot 10^{-5}$	17	$2.43 \cdot 10^{-5}$	First cousins
F09-105 and F10-004	2.14	$3.43 \cdot 10^{-5}$	18	$2.57 \cdot 10^{-5}$	Unrelated
F09-075 and F10-123	2.13	$3.49 \cdot 10^{-5}$	19	$2.72 \cdot 10^{-5}$	Unrelated
F10-069 and F10-122F	2.10	$4.16 \cdot 10^{-5}$	20	$2.86 \cdot 10^{-5}$	Unrelated
F09-040 and F10-026	2.10	$4.16 \cdot 10^{-5}$	21	$3.00 \cdot 10^{-5}$	Unrelated
F10-085 and F10-085F	2.05	$4.96 \cdot 10^{-5}$	22	$3.14 \cdot 10^{-5}$	Unrelated
F09-91F and F10-100	2.00	$6.35 \cdot 10^{-5}$	23	$3.29 \cdot 10^{-5}$	Unrelated
F10-090 andF10-090F	1.95	$8.29 \cdot 10^{-5}$	24	$3.43 \cdot 10^{-5}$	Unrelated
F10-052 and F10-052F	1.92	$9.18 \cdot 10^{-5}$	25	$3.57 \cdot 10^{-5}$	Unrelated
F09-095 and F09-107	1.91	$9.66 \cdot 10^{-5}$	26	$3.72 \cdot 10^{-5}$	Unrelated
F10-106 and F10-106F	1.91	$9.66 \cdot 10^{-5}$	27	$3.86 \cdot 10^{-5}$	Unrelated
F10-084 and F10-084F	1.85	$1.38 \cdot 10^{-4}$	28	$4.00 \cdot 10^{-5}$	Unrelated
F09-125 and F10-119	1.81	$1.75 \cdot 10^{-4}$	29	$4.15 \cdot 10^{-5}$	Unrelated
F10-026 and F10-099	1.75	$2.33 \cdot 10^{-4}$	30	$4.29 \cdot 10^{-5}$	Unrelated
F10-059 and F10-147	1.68	$3.32 \cdot 10^{-4}$	31	$4.43 \cdot 10^{-5}$	Unrelated
F09-065 and F10-004	1.67	$3.47 \cdot 10^{-4}$	32	$4.57 \cdot 10^{-5}$	Unrelated
F10-027 and F10-142	1.64	$3.83 \cdot 10^{-4}$	33	$4.72 \cdot 10^{-5}$	Unrelated
F09-035 and F10-006	1.62	$4.37 \cdot 10^{-4}$	34	$4.86 \cdot 10^{-5}$	Unrelated
F09-103 and F10-017	1.61	$4.43 \cdot 10^{-4}$	35	$5.00 \cdot 10^{-5}$	Unrelated
F10-073 and F10-097	1.60	$4.68 \cdot 10^{-4}$	36	$5.15 \cdot 10^{-5}$	Unrelated
F10-111 and F10-123	1.59	$4.85 \cdot 10^{-4}$	37	$5.29 \cdot 10^{-5}$	Unrelated
F09-044 and F10-062	1.58	$5.21 \cdot 10^{-4}$	38	$5.43 \cdot 10^{-5}$	Unrelated
F09-008 and F09-094	1.56	$5.56 \cdot 10^{-4}$	39	$5.57 \cdot 10^{-5}$	Unrelated
F10-017 and F10-043	1.55	$5.97 \cdot 10^{-4}$	40	$5.72 \cdot 10^{-5}$	Unrelated
F10-037 and F10-37F	1.53	$6.50 \cdot 10^{-4}$	41	$5.86 \cdot 10^{-5}$	Unrelated
F09-040 and F10-135	1.51	$7.05 \cdot 10^{-4}$	42	$6.00 \cdot 10^{-5}$	Unrelated
F09-069 and FF10-006	1.49	$7.84 \cdot 10^{-4}$	43	$6.15 \cdot 10^{-5}$	Unrelated
F09-054 and F10-067	1.48	$8.19 \cdot 10^{-4}$	44	$6.29 \cdot 10^{-5}$	Unrelated
F09-011 and F09-124	1.48	$8.19 \cdot 10^{-4}$	45	$6.43 \cdot 10^{-5}$	Unrelated
F10-034 and F10-115	1.48	$8.22 \cdot 10^{-4}$	46	$6.58 \cdot 10^{-5}$	Unrelated
F10-059 and F10-078	1.47	$8.73 \cdot 10^{-4}$	47	$6.72 \cdot 10^{-5}$	Unrelated
F09-050 and F09-100	1.45	$9.27 \cdot 10^{-4}$	48	$6.86 \cdot 10^{-5}$	Unrelated
F10-026 and F10-113	1.45	$9.34 \cdot 10^{-4}$	49	$7.00 \cdot 10^{-5}$	Unrelated
F09-038 and F09-050	1.44	$9.72 \cdot 10^{-4}$	50	$7.15 \cdot 10^{-5}$	Unrelated

Table 5.11: 33 highest pairwise first cousins LOD scores, mother-foetus pairs not included, and their corresponding Q_r values

Pairs	LOD_{cous}	$\hat{p}$	r	$Q_r = (r/n) \cdot q$	Decision at $q = 0.05$
F10-020 and F10-026	3.28	0.00	1	$1.43 \cdot 10^{-6}$	First cousins
F09-073 and F10-062	3.00	0.00	2	$2.86 \cdot 10^{-6}$	First cousins
F09-047 and F10-079	2.78	$1.27 \cdot 10^{-6}$	3	$4.29 \cdot 10^{-6}$	First cousins
F09-081 and F10-030	2.65	$2.86 \cdot 10^{-6}$	4	$5.72 \cdot 10^{-6}$	First cousins
F09-040 and F10-020	2.59	$3.81 \cdot 10^{-6}$	5	$7.15 \cdot 10^{-6}$	First cousins
F10-089 and F10-140	2.28	$1.91 \cdot 10^{-5}$	6	$8.58 \cdot 10^{-6}$	Unrelated
F09-105 and F10-004	2.14	$3.43 \cdot 10^{-5}$	7	$1.00 \cdot 10^{-5}$	Unrelated
F09-075 and F10-123	2.13	$3.49 \cdot 10^{-5}$	8	$1.14 \cdot 10^{-5}$	Unrelated
F10-069 and F10-122F	2.10	$4.16 \cdot 10^{-5}$	9	$1.29 \cdot 10^{-5}$	Unrelated
F09-040 and F10-026	2.10	$4.16 \cdot 10^{-5}$	10	$1.43 \cdot 10^{-5}$	Unrelated
F09-91F and F10-100	2.00	$6.35 \cdot 10^{-5}$	11	$1.57 \cdot 10^{-5}$	Unrelated
F09-095 and F09-107	1.91	$9.66 \cdot 10^{-5}$	12	$1.72 \cdot 10^{-5}$	Unrelated
F09-125 and F10-119	1.81	$1.75 \cdot 10^{-4}$	13	$1.86 \cdot 10^{-5}$	Unrelated
F10-026 and F10-099	1.75	$2.33 \cdot 10^{-4}$	14	$2.00 \cdot 10^{-5}$	Unrelated
F10-059 and F10-147	1.68	$3.32 \cdot 10^{-4}$	15	$2.15 \cdot 10^{-5}$	Unrelated
F09-065 and F10-004	1.67	$3.47 \cdot 10^{-4}$	16	$2.29 \cdot 10^{-5}$	Unrelated
F10-027 and F10-142	1.64	$3.83 \cdot 10^{-4}$	17	$2.43 \cdot 10^{-5}$	Unrelated
F09-035 and F10-006	1.62	$4.37 \cdot 10^{-4}$	18	$2.57 \cdot 10^{-5}$	Unrelated
F09-103 and F10-017	1.61	$4.43 \cdot 10^{-4}$	19	$2.72 \cdot 10^{-5}$	Unrelated
F10-073 and F10-097	1.60	$4.68 \cdot 10^{-4}$	20	$2.86 \cdot 10^{-5}$	Unrelated
F10-111 and F10-123	1.59	$4.85 \cdot 10^{-4}$	21	$3.00 \cdot 10^{-5}$	Unrelated
F09-044 and F10-062	1.58	$5.21 \cdot 10^{-4}$	22	$3.15 \cdot 10^{-5}$	Unrelated
F09-008 and F09-094	1.56	$5.56 \cdot 10^{-4}$	23	$3.29 \cdot 10^{-5}$	Unrelated
F10-017 and F10-043	1.55	$5.97 \cdot 10^{-4}$	24	$3.43 \cdot 10^{-5}$	Unrelated
F09-040 and F10-135	1.51	$7.05 \cdot 10^{-4}$	25	$3.58 \cdot 10^{-5}$	Unrelated
F09-069 and FF10-006	1.49	$7.84 \cdot 10^{-4}$	26	$3.72 \cdot 10^{-5}$	Unrelated
F09-054 and F10-067	1.48	$8.19 \cdot 10^{-4}$	27	$3.86 \cdot 10^{-5}$	Unrelated
F09-011 and F09-124	1.48	$8.19 \cdot 10^{-4}$	28	$4.00 \cdot 10^{-5}$	Unrelated
F10-034 and F10-115	1.48	$8.22 \cdot 10^{-4}$	29	$4.15 \cdot 10^{-5}$	Unrelated
F10-059 and F10-078	1.47	$8.73 \cdot 10^{-4}$	30	$4.29 \cdot 10^{-5}$	Unrelated
F09-050 and F09-100	1.45	$9.27 \cdot 10^{-4}$	31	$4.43 \cdot 10^{-5}$	Unrelated
F10-026 and F10-113	1.45	$9.34 \cdot 10^{-4}$	32	$4.58 \cdot 10^{-5}$	Unrelated
F09-038 and F09-050	1.44	$9.72 \cdot 10^{-4}$	33	$4.72 \cdot 10^{-5}$	Unrelated

Bonferroni corrected $\hat{p}$ values.
By using the Bonferroni adjustment and putting $\alpha = 0.05$ only nine pairs in the total dataset are concluded as first cousins. Six of those nine are mother-foetus pairs.

FDR

The FDR procedure arranges the $\hat{p}$ values for each LOD score in increasing order $\hat{p}_{(1)} \leq \hat{p}_{(2)} \leq$ ..... $\leq \hat{p}_{(n)}$. The target false discovery rate is $q = 0.05$. Here the result for the non mother-foetus pairs depends on whether the mother-foetus pairs are included in the ranking of the $\hat{p}$-values or not. For that reason the FDR procedure was carried out twice, first including the mother-foetus pairs, see result in table 5.10, and then without including the mother-foetus pairs, see result in table 5.11.

When the mother-foetus pairs are included in the ranking, 17 pairs of individuals are concluded to be first cousins, 11 of them are mother-foetus pairs. Not including the mother-foetus pairs results in one less discoveries of first cousins among the non mother-foetus pairs or five in total. The 95% confidence intervals of $\hat{p}_{(5)}$, (the estimated p-value for the first cousins LOD score of F09-040 and F10-020), and $\hat{p}_{(6)}$, (the estimated p-value for the first cousins LOD score of F10-089 and F10-140), do not collide as can be seen in table 5.8.

5.5 Results

It is evident that the FDR procedure with the false discovery rate fixed at $q = 0.05$ does a better job than the Bonferroni procedure at a significance level of $\alpha = 0.05$ at allocating the mother-foetus pairs in the related group where they should be. The fact the FDR procedure failed to conclude three mother-foetus pairs that had a finite LOD_p score as relatives, gives reason to wonder whether the false discovery rate should be fixed at a higher point than $q = 0.05$ in order to detect all types of first and second order relatives with the half-sibling LOD score. q would have to be raised to 0.42 so that all the mother-foetus pairs with finite LOD_p scores would be concluded as related and as high as 0.69 so that the two mother-foetus pairs with an infinitely negative LOD_p score would be classified as relatives as well.

The FDR procedure correctly concludes all of the 19 mother-foetus pairs with a finite LOD_p score to be a parent and an offspring while the Bonferroni procedure misses three of those 19 pairs. As was expected, the Bonferroni procedure seems to be too strict for this large number of pairwise comparisons and that results in classifying true relatives as unrelated. For this reason conclusions of relatedness will be drawn from the results of the FDR procedure but not the Bonferroni procedure.

Table 5.12 summarizes the pairs that were classified as relatives by the FDR procedure, not including the mother-foetus pairs. The half-sibling LOD score, which Skaug et al (2010) pointed out was a good general test statistic to detect all types of first and second order relatives, detected eight pairs of relatives at $q = 0.05$. The parent-offspring LOD score detected five of those eight pairs as a parent and an offspring. The first cousins LOD score detected five pairs of first cousins but $LOD_{h.sib}$ classified all those pairs as relatives at $q = 0.05$.

In order to come to a conclusion about a specific relatedness for each pair, non genetic evidence, the estimation of their age and age of maturity, has to be taken into account. Recall that fin whales become mature when they're 7-12 years old (Víkingsson, 2005). If the age difference between two fin whales is smaller than the older whale's age of maturity then the conclusion is that it is impossible for them to be a parent and his/her offspring. It is hard to make statements about when fin whales stop being fertile and for that reason there will be no upper limit on the possible age difference between a parent and an offspring in this analysis. The oldest females (they were two) carrying a foetus in this sample were estimated to be 41.5 years old which shows that there are females that are fertile until they reach that age at least.

When the estimated age has been accounted for, the LOD scores for the remaining possible relations have to be compared. $LOD_{h.sib}$, LOD_p and LOD_{cous} compare the probability of the data under the hypothesis of a specific relatedness with the probability of the data under

Table 5.12: Results from the FDR procedure with $q = 0.05$ not including non mother-foetus pairs

Pairs	Related	$LOD_{h.sib}$	Parent-offspring	LOD_p	First cousins	LOD_{cous}
F10-020 and F10-026	Yes	5.07	Yes	7.40	Yes	3.28
F09-073 and F10-062	Yes	4.58	No	$-\infty$	Yes	3.00
F09-081 and F10-030	Yes	4.34	Yes	6.68	Yes	2.65
F09-047 and F10-079	Yes	3.70	No	$-\infty$	Yes	2.78
F09-040 and F10-020	Yes	3.70	No	$-\infty$	Yes	2.59
F10-089 and F10-140	Yes	3.64	Yes	5.47	No	2.28
F09-075 and F10-123	Yes	3.48	Yes	5.38	No	2.13
F09-91F and F10-100	Yes	3.42	Yes	5.43	No	2.00

the null hypothesis of unrelatedness. At this point a comparison of the probability of the data under one specific relatedness with the probability of the data under another specific relatedness is needed. This may be done by simply subtracting one LOD score from the other since $log(\frac{a}{c}) - log(\frac{b}{c}) = log(a) - log(c) - log(b) + log(c) = log(a) - log(b) = log(\frac{a}{b})$.

If the final conclusion for a pair is that they are half-siblings, their estimated age has to be considered again. Recall that it is impossible to distinguish between a pair of half-siblings, grandparent-grandchild pair and uncle/aunt-nephew/niece pair from genetic evidence alone. Information about age can help with that. If the age difference between two fin whales is smaller than the estimated maturity age of the older fin whale plus seven years, the conclusion in this analysis will be that it is impossible for them to be a grandparent and his/her grandchild. There will be no upper limit on the possible age difference between a grandparent and his/her grandchild for the same reason there is no upper limit on the possible age difference between a parent and his/her offspring. Information on the age of the females carrying a foetus in this sample, implies that the age difference between a grandmother and her grandchild can be at least as big as 83 years.

F10-020 and F10-026

The fin whales F10-020, a 39.5 years old female that became mature when she was 10 years old, and F10-026, a 25 years old male, are concluded related from the half-sibling LOD score at $q = 0.05$. They are also concluded as a mother and her son from LOD_p and as first cousins from LOD_{cous}. Since their age difference doesn't exclude that they are a mother and son pair the probability of the data under these specific relatedness hypothesis have to be compared:

$$\begin{aligned}\log(\frac{P(data \mid half - siblings)}{P(data \mid parent\ and\ offspring)}) &= LOD_{h.sib} - LOD_p \\ &= 5.07 - 7.40 = -2.33\end{aligned}$$

F10-020 and F10-026 are more likely to be a parent and an offspring pair than half-siblings.

$$\begin{aligned}log(\frac{P(data \mid parent\ and\ offspring)}{P(data \mid first\ cousins)}) &= LOD_p - LOD_{cous} \\ &= 7.40 - 3.28 = 4.12\end{aligned}$$

F10-020 and F10-026 are more likely to be a parent and her offspring than first cousins. The conclusion is therefore that F10-020 and F10-026 are a mother and her son.

F09-073 and F10-062

The individual whales F09-073, a 14.5 years old male, and F10-062, a 37.5 years old male with an estimated maturity age of 10.5 years, are concluded related from the half-sibling LOD score at $q = 0.05$. They are also concluded as first cousins from LOD_{cous}. The probability of the data under these specific relatedness hypothesis have to be compared:

$$\begin{aligned} log(\frac{P(data \mid half-siblings)}{P(data \mid first\ cousins)}) &= LOD_{h.sib} - LOD_{cous} \\ &= 4.58 - 3.00 = 1.58 \end{aligned}$$

F09-073 and F10-062 are more likely to be half-siblings, a grandfather and his grandson or an uncle and his nephew than to be first cousins. Since their age difference is 22 years (F10-062 was 36.5 years old in 2009) it is impossible to draw further conclusions about their relatedness.

F09-081 and F10-030

The individual whales F09-081, a 15 years old female, and F10-030, a 45 years old female with an estimated maturity age of 11 years, are concluded related from their half-sibling LOD score at $q = 0.05$. They are also concluded as a mother and a daughter from LOD_p and as first cousins from LOD_{cous}. Since their age difference doesn't exclude mother-daughter relations the probability of the data under these three specific relatedness hypothesis have to be compared:

$$\begin{aligned} log(\frac{P(data \mid half-siblings)}{P(data \mid parent\ and\ offspring)}) &= LOD_{h.sib} - LOD_p \\ &= 4.34 - 6.68 = -2.34 \end{aligned}$$

$$\begin{aligned} log(\frac{P(data \mid parent\ and\ offspring)}{P(data \mid first\ cousins)}) &= LOD_p - LOD_{cous} \\ &= 6.68 - 2.65 = 4.03 \end{aligned}$$

The conclusion is that F10-030 and F09-081 are a mother and her daughter.

F09-047 and F10-079

The individual whales F09-047, a 39 years old male with an estimated maturity age of 10 years, and F10-079, a 22 years old male, are concluded related from the half-sibling LOD score at $q = 0.05$. They are also concluded as first cousins from LOD_{cous}. The probability of the data under these specific relatedness hypothesis have to be compared:

$$\begin{aligned} log(\frac{P(data \mid half-siblings)}{P(data \mid first\ cousins)}) &= LOD_{h.sib} - LOD_{cous} \\ &= 3.70 - 2.78 = 0.92 \end{aligned}$$

F09-047 and F10-079 are concluded as half-siblings or a grandfather and his grandson or an uncle and his nephew. Their age difference is 18 years (note that F10-079 was 21 years old in 2009) so no further conclusions can be drawn about their relatedness.

F09-040 and F10-020

The individual whales F09-040, a 20 years old female, and F10-020, a 39.5 years old female with an estimated maturity age of 10 years, are concluded related from the half-sibling LOD score at $q = 0.05$. They are also concluded as first cousins from LOD_{cous}. The probability of the data under these specific relatedness hypothesis have to be compared:

$$log(\frac{P(data \mid half - siblings)}{P(data \mid first\ cousins)}) = LOD_{h.sib} - LOD_{cous}$$
$$= 3.70 - 2.59 = 1.11$$

F10-020 and F09-040 are concluded as half-siblings or a grandmother and her granddaughter or an aunt and her nice. Their age difference is 18.5 years (F10-020 was 38.5 years old in 2009) so no further conclusions can be drawn about their relatedness.

F10-089 and F10-140

The individual whales F10-089, a 47 years old male with an estimated maturity age of 11 years, and F10-140, a 37.5 years old female, are concluded related from the half-sibling LOD score at $q = 0.05$. They are also concluded as a father and his daughter from LOD_p. Since their age difference is less than 11 years then it is impossible for F10-089 to be the father of F10-140.

$$log(\frac{P(data \mid half - siblings)}{P(data \mid parent\ and\ offspring)}) = LOD_{h.sib} - LOD_p$$
$$= 3.64 - 5.47 = -1.83$$

Here it becomes evident how important it is to take non genetic evidence into account but comparison of the LOD scores indicates that these two whales are more likely to be a father and his daughter than half-siblings. Since the estimation of their age indicates that it is impossible, the conclusion is that F10-089 and F10-140 are either half-siblings or an uncle and his niece (or an aunt and her nephew) but grandfather and granddaughter relations can be ruled out because of the age difference.

F09-075 and F10-123

The individual whales F09-075, a 22 years old female with an estimated maturity age of 12 years, and F10-123, a 18.5 years old female, are concluded related from the half-sibling LOD score at $q = 0.05$. They are also concluded as a mother and her daughter from LOD_p. Their age difference is only four and a half year which excludes the probability of them being a mother and her daughter.

$$log(\frac{P(data \mid half - siblings)}{P(data \mid parent\ and\ offspring}) = LOD_{h.sib} - LOD_p$$
$$= 3.48 - 5.38 = -1.9$$

Here, as in the case of F10-089 and F10-140, the genetic evidence indicates that these two fin whales are more likely to be a mother and her daughter than half-siblings. Since the estimation of their age shows that it is impossible, the conclusion is that F09-075 and F10-123 are either half-sisters or an aunt and her niece. Grandmother and granddaughter relations can be ruled out because of the age difference.

F09-091F and F10-100

The foetus F09-091F and F10-100, a 35.5 years old male with a maturity age of 11 years, are concluded related by the half-sibling LOD score at $q = 0.05$. They are also concluded as a parent and an offspring by LOD_p. Since the age of F10-100 doesn't exclude him from being the father of F09-091F the probability of the data under these specific relatedness hypothesis have to be compared:

$$\begin{aligned} log(\frac{P(data \mid half - siblings)}{P(data \mid parent\ and\ offspring}) &= LOD_{h.sib} - LOD_p \\ &= 3.42 - 5.43 = -2.1 \end{aligned}$$

F10-100 and F09-091F are more likely to be a father and his offspring than to be half-siblings. In this case, further auxiliary data is available, the DNA profile of F09-091. A paternity likelihood ratio can be computed for F10-100 and F09-091F since the mother of F09-091F is known. The difference between the parent-offspring likelihood ratio and the paternity likelihood ratio is that the latter one accounts for the mothers profile while the other one doesn't. Balding (2005) provides a good description of the computation of paternity likelihood ratios.

At 13 of 15 loci, the genotype of F09-091 suffices to determine the paternal allele of F09-091F, either because F09-91F is homozygous[1] at those loci or because F09-091F shares exactly one allele with F09-091. At those loci, F10-100 would be excluded from being the father of F09-091F if the paternal allele type was not present in his genotype at one locus or more. That is not the case here, but nothing in the genetic profile of F09-091 excludes F10-100 from being the father of F09-091F. The single locus paternity likelihood ratios for F10-100 and F09-091F, at the loci where the paternal allele is known and F10-100 is homozygous, are computed in the following way:

$$\begin{aligned} LR &= \frac{P(C_p = a_i \mid F = (a_i, a_i)\ is\ the\ father\ of\ C)}{P(C_p = a_i \mid Z\ is\ the\ father\ of\ C)} \\ &= \frac{1}{p(a_i)} \end{aligned} \tag{5.1}$$

C_p stands for the paternal allele of individual C. The numerator is 1 since a father with genotype (a_i, a_i) passes a_i to his offspring with a probability of 1. The denominator is the probability that an allele drawn from Z, some male other than F, is a_i. Since the genetic profile of Z is unavailable, this probability is regarded here as $p(a_i)$, the proportion of a_i alleles in the population of potential fathers. The single locus paternity likelihood ratios for F10-100 and F09-091F, at the loci where the paternal allele is known and F10-100 is heterozygous[2], are computed in the following way:

$$\begin{aligned} LR &= \frac{P(C_p = a_i \mid F = (a_i, a_j)\ is\ the\ father\ of\ C)}{P(C_p = a_i \mid Z\ is\ the\ father\ of\ C)} \\ &= \frac{0.5}{p(a_i)} \\ &= \frac{1}{2 \cdot p(a_i)} \end{aligned} \tag{5.2}$$

[1] Has two copies of the same allele at that locus.
[2] Has two different allele types at that locus.

The numerator is 0.5 sine a father with genotype (a_i, a_j) passes a_i to his offspring with a probability of 0.5. At 2 of 15 loci, F09-091 and F091F have an identical heterozygous genotype while the genotype of F10-100 is homozygous and contains one of F09-091F allele types. At these loci it is not clear which allele of F09-91F is the maternal allele and which one is the paternal allele. The single locus paternity likelihood ratios for F10-100 and F09-091F, at the loci where the paternal allele is unknown and F10-100 is homozygous, are computed in the following way:

$$\begin{aligned} LR &= \frac{P(C = (a_i, a_j) \mid M = (a_i, a_j),\ F = (a_j, a_j)\ is\ the\ father\ of\ C)}{P(C = (a_i, a_j) \mid M = (a_i, a_j),\ Z\ is\ the\ father\ of\ C)} \\ &= \frac{0.5 \cdot 1}{(0.5 \cdot p(a_i) + 0.5 \cdot p(a_j))} \\ &= \frac{1}{p(a_i) + p(a_j)} \end{aligned} \tag{5.3}$$

The mother is denoted here with M. The numerator takes value 0.5 since a father with genotype (a_j, a_j) passes a_j to his offspring with probability 1 while a mother with genotype (a_i, a_j) passes a_i to her offspring with probability 0.5. If F is not the father of C then the two possible transmissions from M to C have to be considered with the proportion of a_i and a_j alleles in the population of potential fathers. That gives the denominator of: $(0.5 \cdot p(a_i) + 0.5 \cdot p(a_j))$ (Balding, 2005).

When the 15 single locus paternity likelihood ratios have been computed for F10-100 and F09-091F, then their paternity LOD score is attained by multiplication and taking the 10th logarithm of the result:

$$\begin{aligned} LOD_{paternity} = \log(&\frac{1}{2 \cdot p_1(159)} \cdot \frac{1}{p_2(193)} \cdot \frac{1}{2 \cdot p_3(125)} \cdot \frac{1}{2 \cdot p_4(125)} \\ &\cdot \frac{1}{2 \cdot p_5(169)} \cdot \frac{1}{p_6(116)} \cdot \frac{1}{p_7(114) + p_7(118)} \cdot \frac{1}{p_8(106) + p_8(112)} \\ &\cdot \frac{1}{2 \cdot p_9(154)} \cdot \frac{1}{2 \cdot p_{10}(215)} \cdot \frac{1}{2 \cdot p_{11}(270)} \cdot \frac{1}{2 \cdot p_{12}(96)} \\ &\cdot \frac{1}{2 \cdot p_{13}(269)} \cdot \frac{1}{2 \cdot p_{14}(207)} \cdot \frac{1}{p_{15}(86)}) \\ = 8.17 \end{aligned}$$

$p_s(a_i)$ is the estimated population allele frequency of a_i at locus s.

The paternity LOD score of F10-100 and F09-091F is higher than their parent-offspring LOD score by: $8.17 - 5.43 = 2.74$. By considering F10-100, F09-091 and F09-091F jointly as a parent-pair and their offspring, F10-100 can now be classified as the father of F09-091F with greater determination than when F10-100 and F09-091F were examined pairwise. The final conclusion here is that F10-100 is the father of F09-091F but that is the same result as Pampoulie et al. (2012) attained when they searched for fathers of the foetuses in this same sample.

Table 5.13: DNA profiles of F09-091, F09-091F and F10-100

Locus	F09-091	F09-91F	F10-100
EV001	157/163	159/163	159/171
EV037	193/193	193/193	193/193
GT011	127/131	125/131	117/125
GT023	127/129	125/129	125/129
GT195	161/175	161/169	169/173
GT211	120/120	116/120	116/116
GT271	114/118	114/118	118/118
GT310	106/112	106/112	112/112
GT575	154/156	154/154	152/154
GATA028	199/219	199/215	215/227
GATA053	262/262	262/270	258/270
GATA098	100/100	96/100	96/108
GATA417	269/277	269/269	269/285
GGAA520	201/223	207/223	207/219
TAA023	86/86	86/86	86/86

Table 5.14: Detected pairs of relatives within the sample

Pairs	Conclusion
F10-020 and F10-026	Mother and son
F10-062 and F09-073	Half-brothers/grandfather and grandson/uncle and nephew
F10-030 and F09-081	Mother and daughter
F09-047 and F10-079	Half-brothers/grandfather and grandson/uncle and nephew
F10-020 and F09-040	Half-sisters/grandmother and granddaughter/aunt and niece
F10-089 and F10-140	Half-siblings/uncle and niece/aunt and nephew
F09-075 and F10-123	Half-sisters/niece and aunt
F10-100 and F09-91F	Father and offspring

The final result of the analysis has been summarized in table 5.14. The test procedure detected all in all eight pairs of related individuals within the dataset of 34 959 pairs. Three of those related pairs were classified as a parent and his/her offspring, three were classified as half-siblings or a grandparent-grandchild pair or an uncle/aunt-nephew/niece pair and two were classified as half-siblings or an uncle/aunt-niece/nephew pair.

Chapter 6

Discussions

Very little is known about the second biggest marine mammal in the world, the fin whale. Fin whales are very difficult to observe so uncertainties remain about their genetic structure, abundance, mating strategies and migration patterns (Ægisson and Hlíðberg, 2010, Víkingsson, 2005, Pampoulie et al., 2012). Several genetic studies of this species have been performed over the last few decades in order to find out more about their migration patterns but the results were so far inconclusive. Sighting surveys indicate that fin whales are most commonly seen alone or in pairs and relatedness analyses confirmed that related individuals more commonly occur at the same feeding location (Pampoulie et al., 2012).

Iceland has maintained an individual-based DNA-registry for fin whales for some time. The present study used data from this registry by using a general statistical procedure for detecting pairs of relatives. Three types of relatedness were of interest, half-siblings, parent-offspring and first cousins relations. Relatedness was tested among 265 individuals which means that $\frac{265 \cdot 264}{2} = 34\,980$ pairwise relations were examined. 21 known mother-foetus pairs were present in the sample. These pairs were very beneficial for the analysis since assumptions about the quality of the test procedure could be drawn from the ability to detect their relatedness. Hardy-Weinberg and linkage equilibrium were assumed and the population allele frequencies were estimated directly from the dataset excluding the foetuses. Detection of relatives was done by computing pairwise LOD scores for the 265 individuals in the sample for each relatedness of interest. If D_i and D_j were the genetic profiles of individuals i and j than the LOD score for their relatedness would be denoted with (Skaug et al., 2010):

$$LOD_{i,j} = \log(\frac{P(D_i, D_j \mid H_1 : related)}{P(D_i, D_j \mid H_0 : unrelated)})$$

A high LOD score indicates relatedness but that entails an issue of what should be considered to be high enough. That issue was accounted for by reporting a single p-value with each LOD score. The p-values were attained via simulation. 265 unrelated individuals were simulated with the same population allele frequencies as the ones estimated from the dataset and then their pairwise LOD scores were computed. This procedure was replicated at least 60 times. The p-values were computed by comparing the original LOD scores with the simulated ones but $p_{i,j}$ can be described as the probability of attaining as extreme or more extreme LOD score than $LOD_{i,j}$ just by chance. All computations and simulations were done by using the open source program R (R Development Core Team, 2011), and the codes written can be

found in appendix D.

Relatedness was tested for every possible pair in the dataset. The high number of pairwise comparisons raised a well known statistical issue, the problem of multiple testing. The problem of multiple testing was addressed by comparing two multiple adjustment methods, the Bonferroni correction and the FDR procedure. The Bonferroni correction controls the family wise error rate while the FDR procedure controls the false discovery rate. The Bonferroni correction is known to be very conservative when the number of multiple test is high. The FDR procedure is more flexible since it takes the number of erroneous false discoveries of relatedness into account instead of only the question of whether any error was made (Benjamini and Hochberg, 1995).

In this study, conclusions about relatedness are drawn from the result of the FDR procedure since it performed better than the Bonferroni correction at allocating the mother-foetus pairs in the related group. The FDR procedure, with the false discovery rate fixed at $q = 0.05$, correctly concluded all of the 19 mother-foetus pairs with a finite LOD_p score[1] to be a mother and her offspring while the Bonferroni procedure, at a significance level of $\alpha = 0.05$, missed three of those 19 pairs. As was expected, the Bonferroni procedure was too strict for this large number of simultaneous pairwise comparisons with the cost of not detecting true relatives.

At $q = 0.05$, eight pairs of relatives were detected in the sample[2]. Three of those pairs were classified as a parent and an offspring pair, three pairs were classified as half-siblings or a grandparent-grandchild pair or an uncle/aunt-nephew/niece pair and the remaining two of those eight pairs were classified as half-siblings or an uncle/aunt-nephew/nice pair. No first cousins were detected within the dataset. The result might have been different if the genetic profiles contained information about more loci. There is of course a possibility, that all the eight matches of relatives were incidental and due to low number of loci employed but there's also a possibility that information about more loci would have resulted in an increased rate of detected relatives.

In their paper 'Detecting dyads of related individuals in large collections of DNA-profiles by controlling the false discovery rate', Skaug et al. (2010) considered 'half-siblings' to be a reasonable choice for a general test to detect all types of close 1st- and 2nd order relationships. The results of this study are some what in harmony with that recommendation. At $q = 0.05$ the half-sibling LOD score detected all the pairs that LOD_p concluded as a parent and an offspring as well as all the pairs that LOD_{cous} classified as first cousins. However $LOD_{h.sib}$ failed to detect five mother-foetus pairs as relatives, thereof three with a finite LOD_p score. That indicates that, in order for the half-sibling LOD score to detect all relatives of $1st$-order, the false discovery rate has to be fixed at a higher level than $q = 0.05$. q had to be raised to 0.42 so that $LOD_{h.sib}$ would have detected all mother-foetus pairs with a finite LOD_p and as high as 0.69 so that the two mother-foetus pairs with an infinitely negative LOD_p score would have been classified as relatives as well.

[1] The mother-foetus pairs are 21 in the sample but two of those pairs don't have matching alleles at one locus due to mutation or a typing error which results in an infinitely negative parent-child LOD score.

[2] Mother-foetus pairs excluded.

In 'A note on a mother-foetus pair and alleged father match in the Atlantic fin whale (*Balaenoptera physalus*) off Iceland' Pampoulie et al. (2012) analysed the same Icelandic fin whale registry that is used in the present study. They did so by statistically comparing the genotype profiles of the 23 mother-foetus pairs to that of the 139 potential fathers within the database. The software WHICHPARENTS (available at: *http://www-bml.ucdavis.edu/whichparents.html*) was used to assess potential crosses among mother-foetus and alleged father, using 0-4 potential misses. The exclusion program WHICHPARENTS revealed the presence of one possible cross between a mother-foetus pair and an alleged father when run with 0 miss procedure, i.e. a 100% match. Additional analyses of this possible family, involving the mother F09-091, her foetus F09-091F and the alleged father F10-100, was performed in the software PATCAN v1.2 (available on request to J.A. Riancho; Riancho and Zarrabeitia (2003)) to assess the paternity probability of the potential father. It revealed a high likelihood and probability associated with the hypothesis that the alleged father was the true biological father of the foetus F09-091F. Pampoulie et al. point out that their match might be incidental and due to the low number of loci employed. They argue that at least two hypothesis can be considered to explain the observed trio-match:

1. The detected mating pair occurring at the same mating location exhibited a similar migration habit during the winter.

2. The detected mating pair may originally belong to two different populations (or mating locations) among which gene flow may not be restricted, which might indicate that individual fin whale from different mating location may roam across the North Atlantic during the winter feeding migration.

F10-100, F09-091 and F09-091F were also classified as a parent-pair and their offspring in the present analysis. $LOD_{h.sib}$ detected F10-100 and F09-091F as relatives at $q = 0.05$ and LOD_p classified them as a parent and an offspring. After the profiles of F10-100, F09-091 and F09-091F had been examined jointly by computing the paternity LOD score for F10-100 and F09-091F, this trio was concluded as a parent-pair and their offspring.

The aim of relatedness detection studies varies. Here, the main interest was the performance of the statistical procedure. The mother-foetus pairs within the Icelandic fin whale registry were extremely valuable from that perspective. The test procedures seemed to operate well at detecting relatives and classifying their relatedness. It's ability to detect the mother-foetus pairs as a parent and offspring was very good, and it's ability to conclude them as related, by using the half-sibling LOD score, was also good if one would be content with a high false discovery rate.

Of course, the assumptions made, in the process of designing the test procedure, limit its performance, but no statistical test can take into account the complexity of organims-life cycle. Hardy-Weinberg and linkage equilibrium are assumed and, since the true population allele frequencies of fin whales in the EGI area are unknown, the allele frequencies were estimated directly from the dataset. Another limitation of the test procedure, since it is based on pairwise comparisons, is that it only takes two individuals into account at a time and that can lead to inconsistent pedigree results. For example, it is possible, in the case of simultaneous pairwise comparisons, that individuals A and B are classified as full siblings and that B and C

are classified as full sibling but at the same time A and C are classified as half-siblings[3] (Fernández and Toro, 2006). Also in parentage analysis, it is possible that A and C are classified as a father and his offspring and that B and C are classified as a mother and her offspring in pairwise comparisons, but, when considered jointly, this family might be incompatible with a parent-pair and offspring relationship (Jones and Wang, 2009).

The present analysis had very few matches of relatives so it was pretty straight forward to check for inconsistent pedigree results. F10-020 has two detected relatives within the database. She was classified as the mother of F10-026 as well as a half-sister/grandmother/aunt of F09-040 but those relations do obviously not result in an inconsistent pedigree. F10-100 was detected as the father of the foetus F09-91F. In that case, F10-100, F09-091 and F09-091F had to been considered jointly as a family. That was done by computing the paternity LOD score for F10-100 and F09-091F which revealed that the trio was compatible with a parent-pair and offspring relationship. In some cases it might be more difficult to check for inconsistent results, such as in the case of studies that result in a high number of detected dyads of relatives. It might then be more suitable to use an alternative computer program like COLONY (available at: https://www.zsl.org/science/research-projects/software/colony,1154,AR.html) that implements full-pedigree likelihood methods to simultaneously infer sibship and parentage among individuals, with likelihood considered over the entire pedigree configuration (Jones and Wang, 2009).

There are many possibilities for further work with the present procedure. Regarding the fin whale data, it would be interesting to test if there are any full siblings within the dataset. Discovery of first cousins would have implied that there are full sibling fin whales out there (since first cousins are children of full siblings), but no first cousins were detected within the present dataset. The study could be used to evaluate ecological information of fin whales in Icelandic waters. For example, the discoveries of relatives could be regarded as a mark-recapture experiment and used for abundance estimation (Skaug and Oien, 2005). The codes given in appendix D can be used for testing for relatedness within other genetic datasets, the only requirement is that the genetic data is on matrix form. Developing the codes into a more user-friendly mode, for example as a R package, might be of interest for relatedness analysis of research focusing on mark-recapture genetic studies using non-lethal techniques (biopsy). If one were to use the statistical procedure, presented in this paper, to detect many different types of relatives within a large database, the following steps are recommended:

1. Compute all pairwise half-sibling LOD scores
2. Divide the pairs into two groups, 'Related group of pairs' and 'Unrelated group of pairs' based on how the FDR procedure classifies the pairs at a rather high q, for example $q = 0.5$
3. The 'Related group of pairs' is searched for parent-offspring pairs, first cousins, siblings and half-siblings at a lower q

Since the 'Related group of pairs' should be considerably smaller than the original group of pairs, following those steps, instead of computing different LOD for each relatedness hypothesis, could save a lot of computing time in the case of very large datasets.

[3]It became clear in the explanatory example in chapter 4.3 that the test procedure concludes full siblings to be half-siblings if they don't have an identical genotype at any locus.

Bibliography

Ægisson, S. and Hlíðberg, J.B. *Hvalir*. JPV-Forlagid, 2010.

Balding, D.J. *Weight-of-evidence for Forensic DNA Profiles*. John Wiley & Sons Hoboken, NJ, 2005.

Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

Bérubé, M. and Palsbøll, P.J. Identification of sex in cetaceans by multiplexing with three zfx and zfy specific primers. *Molecular Ecology*, 5(2):283–287, 1996.

Casella, G. and Berger, R. *Statistical Inference*. Duxbury, 2 edition, 2002.

Dorai-Raj, S. *binom: Binomial Confidence Intervals For Several Parameterizations*, 2009. URL `http://CRAN.R-project.org/package=binom`. R package version 1.0-5.

Dyer, R.J. *gstudio: GeneticStudio packages for R.*, 2012. URL `http://CRAN.R-project.org/package=gstudio`. R package version 0.8.

Fernández, J. and Toro, MA. A new method to estimate relatedness from molecular markers. *Molecular Ecology*, 15(6):1657–1667, 2006.

Hartl, D.L. and Jones, E.W. *Genetics: Principles and Analysis*. Jones and Bartlett publishers, 4 edition, 1998.

Huber, W.; von Heydeberck, A., and Vingron, M. *Handbook of Statistical Genetics*, volume 1, chapter Analysis of Microarray Gene Expression Data, pages 203–230. Wiley-Interscience, 3 edition, 2007.

Johnson, R.A. and Wichern, D.W. *Applied Multivariate Statistical Analysis*. Pearson Education International, 6 edition, 2007.

Jones, O.R. and Wang, J. Colony: a program for parentage and sibship inference from multilocus genotype data. *Molecular Ecology Resources*, 10(3):551–555, 2009.

Nielsen, R.; Mattila, D.K.; Clapham, P.J., and Palsbøll, P.J. Statistical approaches to paternity analysis in natural populations and applications to the north atlantic humpback whale. *Genetics*, 157(4):1673–1682, 2001.

Pampoulie, C.; Ólafsdóttir, G.; Hauksdóttir, S.; Skírnisdóttir, S.; Ólafsson, K.; Magnúsdóttir, S.; Chosson, V.; Halldórsson, S.D.; Ólafsdóddir, D.; Gunnlaugsson, T.; Daníelsdóttir, A.K., and Víkingsson, G.A. A note on a mother-foetus pair and alleged father match in the atlantic fin whale (*Balaenoptera physalus*) off iceland. *Journal of Cetacean Research and Management*, 2012.

Pounds, S.B.; Cheng, C., and Onar, A. *Handbook of Statistical Genetics*, volume 1, chapter Statistical Inference for Microarray Studies, pages 231–266. Wiley-Interscience, 3 edition, 2007.

R Development Core Team, . *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

Riancho, J.A. and Zarrabeitia, M.T. A windows-based software for common paternity and sibling analyses. *Forensic science international*, 135(3):232–234, 2003.

Rizzo, M.L. *Statistical computing with R*. Chapman & Hall/CRC, 2007.

Russell, J.C.; Abdelkrim, J., and Fewster, R.M. Early colonisation population structure of a norway rat island invasion. *Biological Invasions*, 11(7):1557–1567, 2009.

Sigurjónsson, J. *Villt íslensk spendýr*, chapter Hvalrannsóknir við Ísland, pages 103–146. Hið Íslenska náttúrufræðifélag-Landvernd, 1993.

Sigurðsson, Þ. and Magnússon, Á., editors. volume 163. Marine Research Institute of Iceland, 2012.

Skaug, H.J. Allele-sharing methods for estimation of population size. *Biometrics*, 57(3): 750–756, 2001.

Skaug, H.J. and Oien, N. Genetic tagging of male north atlantic minke whales through comparison of maternal and foetal dna-profiles. *Journal of Cetacean Research and Management*, 7(2):113–117, 2005.

Skaug, H.J.; Berube, M., and Palsbøll, P.J. Detecting dyads of related individuals in large collections of dna-profiles by controlling the false discovery rate. *Molecular Ecology Resources*, 10(4):693–700, 2010.

Speed, T.P. and Zhao, H. *Handbook of Statistical Genetics*, volume 1, chapter Chromosome Maps, pages 3–39. Wiley-Interscience, 3 edition, 2007.

Thompson, E.A. *Pedigree Analysis in Human Genetics*. Johns Hopkins University Press Baltimore, 1986.

Víkingsson, G.A. *Íslensk spendýr*, chapter Langreyður, pages 204–211. Vaka-Helgafell, 2005.

Weir, B.S. *Handbook of Statistical Genetics*, volume 2, chapter Forensics, pages 1368–1390. Wiley-Interscience, 3 edition, 2007.

Appendices

Appendix A

Kinship coefficients

One of the simplest probabilities of gene identity by descent is the classical kinship coefficient. The kinship coefficient, k_j, between two individuals is defined as the probability that they have inherited j alleles at a locus identical by descent given a certain relatedness.

A.1 Siblings

Full siblings can share 0, 1 or 2 alleles at a locus but the probabilities differ. In order to find the appropriate relatedness coefficients for siblings the following question has to be answered: *For $j = 0, 1, 2$, given that two individuals are siblings, whose parents have alleles (a, b) and (c, d) at a given locus, what is the probability that they have inherited j alleles IBD at that locus?*

$$\begin{aligned} k_0 &= P(0 - ibd \mid siblings) \\ &= P(ind_i = (a,c) \bigcap ind_j = (b,d) \mid siblings) \\ &+ P(ind_i = (b,d) \bigcap ind_j = (a,c) \mid siblings) \\ &+ P(ind_i = (a,d) \bigcap ind_j = (b,c) \mid siblings) \\ &+ P(ind_i = (b,c) \bigcap ind_j = (a,d) \mid siblings) \\ &= 2 \cdot (\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2}) + 2 \cdot (\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2}) \\ &= \frac{4}{16} = \frac{1}{4} \end{aligned} \tag{A.1}$$

$$
\begin{aligned}
k_1 &= P(1-ibd \mid siblings) \\
&= P(ind_i = (a,c) \bigcap ind_j = (a,d) \mid siblings) \\
&+ P(ind_i = (a,d) \bigcap ind_j = (a,c) \mid siblings) \\
&+ P(ind_i = (b,c) \bigcap ind_j = (b,d) \mid siblings) \\
&+ P(ind_i = (b,d) \bigcap ind_j = (b,c) \mid siblings) \\
&+ P(ind_i = (c,b) \bigcap ind_j = (c,a) \mid siblings) \\
&+ P(ind_i = (c,a) \bigcap ind_j = (c,b) \mid siblings) \\
&+ P(ind_i = (d,a) \bigcap ind_j = (d,b) \mid siblings) \\
&+ P(ind_i = (d,b) \bigcap ind_j = (d,a) \mid siblings) \\
&= 2 \cdot (\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2}) + 2 \cdot (\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2}) \\
&+ 2 \cdot (\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2}) + 2 \cdot (\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2}) \\
&= \frac{8}{16} = \frac{1}{2}
\end{aligned}
\tag{A.2}
$$

$$
\begin{aligned}
k_2 &= P(2-ibd \mid siblings) \\
&= P(ind_i = (a,c) \bigcap ind_j = (a,c) \mid siblings) \\
&+ P(ind_i = (a,d) \bigcap ind_j = (a,d) \mid siblings) \\
&+ P(ind_i = (b,c) \bigcap ind_j = (b,c) \mid siblings) \\
&+ P(ind_i = (b,d) \bigcap ind_j = (b,d) \mid siblings) \\
&= \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \\
&+ \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \\
&= \frac{4}{16} = \frac{1}{4}
\end{aligned}
\tag{A.3}
$$

A.2 Half-Siblings

Half-siblings[1] have one parent in common and can therefore inherit 0 or 1 alleles IBD but never 2.

For $j = 0, 1, 2$, given that two individuals are half-siblings, whose common parent has alleles (a, b) at a locus, what is the probability that they have inherited j alleles IBD at that locus?

[1] The kinship coefficients for grandparent-grandchild relations and uncle/aunt-nephew/niece relations are the same as for half-siblings.

$$
\begin{aligned}
k_0 &= P(0-ibd \mid half\ siblings) \\
&= P(ind_i = (a) \bigcap ind_j = (b) \mid half - siblings) \\
&+ P(ind_i = (b) \bigcap ind_j = (a) \mid half - siblings) \\
&= \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2}
\end{aligned}
\tag{A.4}
$$

$$
\begin{aligned}
k_1 &= P(1-ibd \mid half\ siblings) \\
&= P(ind_i = (a) \bigcap ind_j = (a) \mid half - siblings) \\
&+ P(ind_i = (b) \bigcap ind_j = (b) \mid half - siblings) \\
&= \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2}
\end{aligned}
\tag{A.5}
$$

$$
k_2 = P(2-ibd \mid half - siblings) = 0 \tag{A.6}
$$

A.3 First Cousins

First cousins have one set of grand-parents pair in common. They can therefore inherit 0 or 1 alleles IBD at a locus but never 2 (assuming there is no inbreeding). The kinship coefficients are attained by answering the following question:
For $j = 0, 1, 2$, given that two individuals are first cousins, whose common pair of grand-parents have alleles (a, b) and (c, d) at a locus, what is the probability that they have inherited j alleles IBD at that locus?

$$
\begin{aligned}
k_0 &= P(0-ibd \mid 1st.cousins) \\
&= P(ind_i = (a) \bigcap ind_j = (b) \mid 1st.cousins) \\
&+ P(ind_i = (b) \bigcap ind_j = (a) \mid 1st.cousins) \\
&+ P(ind_i = (a) \bigcap ind_j = (c) \mid 1st.cousins) \\
&+ P(ind_i = (c) \bigcap ind_j = (a) \mid 1st.cousins) \\
&+ P(ind_i = (a) \bigcap ind_j = (d) \mid 1st.cousins) \\
&+ P(ind_i = (d) \bigcap ind_j = (a) \mid 1st.cousins) \\
&+ P(ind_i = (b) \bigcap ind_j = (c) \mid 1st.cousins) \\
&+ P(ind_i = (c) \bigcap ind_j = (b) \mid 1st.cousins) \\
&+ P(ind_i = (b) \bigcap ind_j = (d) \mid 1st.cousins) \\
&+ P(ind_i = (d) \bigcap ind_j = (b) \mid 1st.cousins) \\
&+ P(ind_i = (c) \bigcap ind_j = (d) \mid 1st.cousins) \\
&+ P(ind_i = (d) \bigcap ind_j = (c) \mid 1st.cousins) \\
&= 2 \cdot \frac{1}{4} \cdot \frac{1}{4} + 2 \cdot \frac{1}{4} \cdot \frac{1}{4} + 2 \cdot \frac{1}{4} \cdot \frac{1}{4} \\
&+ 2 \cdot \frac{1}{4} \cdot \frac{1}{4} + 2 \cdot \frac{1}{4} \cdot \frac{1}{4} + 2 \cdot \frac{1}{4} \cdot \frac{1}{4} \\
&= \frac{12}{16} = \frac{3}{4}
\end{aligned}
\tag{A.7}
$$

$$
\begin{aligned}
k_1 &= P(1-ibd \mid 1st.cousins) \\
&= P(ind_i = (a) \bigcap ind_j = (a) \mid 1st.cousins) \\
&+ P(ind_i = (b) \bigcap ind_j = (b) \mid 1st.cousins) \\
&+ P(ind_i = (c) \bigcap ind_j = (c) \mid 1st.cousins) \\
&+ P(ind_i = (d) \bigcap ind_j = (d) \mid 1st.cousins) \\
&= \frac{1}{4} \cdot \frac{1}{4} + \frac{1}{4} \cdot \frac{1}{4} + \frac{1}{4} \cdot \frac{1}{4} + \frac{1}{4} \cdot \frac{1}{4} \\
&= \frac{4}{16} = \frac{1}{4}
\end{aligned}
\tag{A.8}
$$

$$
k_2 = P(2-ibd \mid 1st.cousins) = 0 \tag{A.9}
$$

Appendix B

Relatedness Likelihood Ratios

Consider allele information of individuals i and j at S loci. Under the assumption of linkage equilibrium it is possible to test their relatedness by computing the likelihood ratio for each locus separately and then multiply those ratios together and take the logarithm to get the LOD score.
$LR_{i,j}(s) = \frac{P(D_{i,s}, D_{i,s}|H_1)}{P(D_{i,s}, D_{j,s}|H_0)}$
is the likelihood ratio at locus s. The computations in the present study are done under the assumption of Hardy-Weinberg- and linkage equilibrium. At a single locus, the probability of two microsatellite based DNA-profiles, D_i and D_j, given the null hypothesis of unrelatedness, is:
$P(D_i, D_j \mid unrelated) = p(a_i^{(1)}) \cdot p(a_i^{(2)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)})$
where $(a_{i,s}^{(1)}, a_{i,s}^{(2)})$ is the genotype of individual i at locus s and $p(a_{i,s}^{(m)})$ is the population frequency for whatever type allele $a_{i,s}^{(m)}$ is with $m = 1, 2$. This probability is used in the following computation of likelihood ratios.

B.1 Full Siblings Likelihood Ratio at a Single Locus

H_0 : Individual i and j are siblings.
H_1 : Individual i and j are unrelated

There are three random events that need conditioning:

1. How many alleles are shared IBD: 0,1 or 2.
2. Which allele was inherited from the mother, which allele was inherited from the father.
3. Given 2), did individuals i and j inherit the same allele from the same parent.

$$
\begin{aligned}
P(D_i, D_j \mid siblings) = & \frac{1}{1} \cdot k_0(siblings) \cdot p(a_i^{(1)}) \cdot p(a_i^{(2)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \\
& + \frac{1}{4} \cdot k_1(siblings) \\
& \cdot (p(a_i^{(2)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(1)} = a_j^{(1)}) \\
& + (p(a_i^{(1)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(2)} = a_j^{(1)}) \\
& + (p(a_i^{(2)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(1)} = a_j^{(2)}) \\
& + (p(a_i^{(1)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(2)} = a_j^{(2)})) \\
& + \frac{1}{2} \cdot k_2(siblings) \\
& \cdot (p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(1)} = a_j^{(1)} \bigcap a_i^{(2)} = a_j^{(2)}) \\
& + p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(1)} = a_j^{(2)} \bigcap a_i^{(2)} = a_j^{(1)})) \\
& = \frac{1}{1} \cdot \frac{1}{4} \cdot p(a_i^{(1)}) \cdot p(a_i^{(2)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \\
& + \frac{1}{4} \cdot \frac{1}{2} \\
& \cdot (p(a_i^{(2)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(1)} = a_j^{(1)}) \\
& + (p(a_i^{(1)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(2)} = a_j^{(1)}) \\
& + (p(a_i^{(2)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(1)} = a_j^{(2)}) \\
& + (p(a_i^{(1)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(2)} = a_j^{(2)})) \\
& + \frac{1}{2} \cdot \frac{1}{4} \\
& \cdot (p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(1)} = a_j^{(1)} \bigcap a_i^{(2)} = a_j^{(2)}) \\
& + p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(1)} = a_j^{(2)} \bigcap a_i^{(2)} = a_j^{(1)}))
\end{aligned}
\tag{B.1}
$$

$$
\begin{aligned}
LR_{sib} = & \frac{P(D_i, D_j \mid siblings)}{P(D_i, D_j \mid unrelated)} \\
= & \frac{1}{4} \\
& + \frac{1}{2} \cdot (\frac{I(a_i^{(1)} = a_j^{(1)}) + I(a_i^{(1)} = a_j^{(2)})}{4 \cdot p(a_i^{(1)})} + \frac{I(a_i^{(2)} = a_j^{(1)}) + I(a_i^{(2)} = a_j^{(2)})}{4 \cdot p(a_i^{(2)})}) \\
& + \frac{1}{4} \cdot (\frac{I(a_i^{(1)} = a_j^{(1)} \bigcap a_i^{(2)} = a_j^{(2)}) + I(a_i^{(1)} = a_j^{(2)} \bigcap a_i^{(2)} = a_j^{(1)}))}{2 \cdot p(a_i^{(1)}) \cdot p(a_i^{(2)})}) \\
= & \frac{1}{4} + \frac{1}{2} \cdot LR_p + \frac{1}{4} \cdot LR_{id}
\end{aligned}
\tag{B.2}
$$

B.2 Half-Siblings Likelihood Ratio at a Single Locus

H_0 : Individual i and j are half-siblings
H_1 : Individual i and j are unrelated

There are three random events that need conditioning:

1. How many alleles are shared IBD: 0 or 1.
2. Which allele was inherited from the shared parent.
3. Given 2), did individuals i and j inherit the same allele from their shared parent.

$$
\begin{aligned}
P(D_i, D_j \mid half\ siblings) = & \frac{1}{1} \cdot k_0(half - siblings) \cdot p(a_i^{(1)}) \cdot p(a_i^{(2)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \\
& + \frac{1}{4} \cdot k_1(half - siblings) \\
& \cdot (p(a_i^{(2)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(1)} = a_j^{(1)}) \\
& + (p(a_i^{(1)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(2)} = a_j^{(1)}) \\
& + (p(a_i^{(2)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(1)} = a_j^{(2)}) \\
& + (p(a_i^{(1)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(2)} = a_j^{(2)})) \\
& + \frac{1}{2} \cdot k_2(half - siblings) \\
& \cdot (p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(1)} = a_j^{(1)} \bigcap a_i^{(2)} = a_j^{(2)}) \\
& + p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(1)} = a_j^{(2)} \bigcap a_i^{(2)} = a_j^{(1)})) \\
& = \frac{1}{1} \cdot \frac{1}{2} \cdot p(a_i^{(1)}) \cdot p(a_i^{(2)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \\
& + \frac{1}{4} \cdot \frac{1}{2} \\
& \cdot (p(a_i^{(2)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(1)} = a_j^{(1)}) \\
& + (p(a_i^{(1)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(2)} = a_j^{(1)}) \\
& + (p(a_i^{(2)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(1)} = a_j^{(2)}) \\
& + (p(a_i^{(1)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(2)} = a_j^{(2)})) \\
& + \frac{1}{2} \cdot 0 \\
& \cdot (p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(1)} = a_j^{(1)} \bigcap a_i^{(2)} = a_j^{(2)}) \\
& + p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(1)} = a_j^{(2)} \bigcap a_i^{(2)} = a_j^{(1)}))
\end{aligned}
\tag{B.3}
$$

$$
\begin{aligned}
LR_{h.sib} &= \frac{P(D_i, D_j \mid half\ siblings)}{P(D_i, D_j \mid unrelated)} \\
&= \frac{1}{2} \\
&+ \frac{1}{2} \cdot \left(\frac{I(a_i^{(1)} = a_j^{(1)}) + I(a_i^{(1)} = a_j^{(2)})}{4 \cdot p(a_i^{(1)})} + \frac{I(a_i^{(2)} = a_j^{(1)}) + I(a_i^{(2)} = a_j^{(2)})}{4 \cdot p(a_i^{(2)})}\right) \\
&= \frac{1}{2} + \frac{1}{2} \cdot LR_p + 0 \cdot LR_{id}
\end{aligned}
\tag{B.4}
$$

B.3 First Cousins Likelihood Ratio at a Single Locus

H_0 : Individual i and j are first cousins.
H_1 : Individual i and j are unrelated

There are three random events that need conditioning:

1. How many alleles are shared IBD: 0 or 1.
2. Which allele was inherited from their shared pair of grand-parents.
3. Given 2), did individual i and j inherit the same allele from their shared pair of grand-parents.

$$
\begin{aligned}
P(D_i, D_j \mid first\ cousins) = & \frac{1}{1} \cdot k_0(first\ cousins) \cdot p(a_i^{(1)}) \cdot p(a_i^{(2)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \\
& + \frac{1}{4} \cdot k_1(first\ cousins) \\
& \cdot (p(a_i^{(2)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(1)} = a_j^{(1)}) \\
& + (p(a_i^{(1)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(2)} = a_j^{(1)}) \\
& + (p(a_i^{(2)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(1)} = a_j^{(2)}) \\
& + (p(a_i^{(1)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(2)} = a_j^{(2)})) \\
& + \frac{1}{2} \cdot k_2(first\ cousins) \\
& \cdot (p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(1)} = a_j^{(1)} \bigcap a_i^{(2)} = a_j^{(2)}) \\
& + p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(1)} = a_j^{(2)} \bigcap a_i^{(2)} = a_j^{(1)})) \\
& = \frac{1}{1} \cdot \frac{3}{4} \cdot p(a_i^{(1)}) \cdot p(a_i^{(2)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \\
& + \frac{1}{4} \cdot \frac{1}{4} \\
& \cdot (p(a_i^{(2)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(1)} = a_j^{(1)}) \\
& + (p(a_i^{(1)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(2)} = a_j^{(1)}) \\
& + (p(a_i^{(2)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(1)} = a_j^{(2)}) \\
& + (p(a_i^{(1)}) \cdot p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(2)} = a_j^{(2)})) \\
& + \frac{1}{2} \cdot 0 \\
& \cdot (p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(1)} = a_j^{(1)} \bigcap a_i^{(2)} = a_j^{(2)}) \\
& + p(a_j^{(1)}) \cdot p(a_j^{(2)}) \cdot I(a_i^{(1)} = a_j^{(2)} \bigcap a_i^{(2)} = a_j^{(1)}))
\end{aligned}
\tag{B.5}
$$

$$
\begin{aligned}
LR_{cous} = & \frac{P(D_i, D_j \mid first\ cousins)}{P(D_i, D_j \mid unrelated)} \\
= & \frac{3}{4} \\
& + \frac{1}{4} \cdot (\frac{I(a_i^{(1)} = a_j^{(1)}) + I(a_i^{(1)} = a_j^{(2)})}{4 \cdot p(a_i^{(1)})} + \frac{I(a_i^{(2)} = a_j^{(1)}) + I(a_i^{(2)} = a_j^{(2)})}{4 \cdot p(a_i^{(2)})}) \\
= & \frac{3}{4} + \frac{1}{4} \cdot LR_p + 0 \cdot LR_{id}
\end{aligned}
\tag{B.6}
$$

Appendix C

R Codes for the Explanatory Example

This sections includes the codes that were used in the explanatory example in chapter 4.3. All computations were done by using the open source program R, version 2.14.1, (R Development Core Team, 2011).

C.1 Simulation of Individuals

The three individuals were simulated by the following code written by Hans Julius Skaug. Two of them were simulated to be full siblings while the third one was simulated to be unrelated to them. Here individuals A, B and C are referred to as individuals 1, 2 and 3.

```
S<-NULL
S=c(9,18,20,14,10,12,12,11,17,16) # Number of alleles per locus i.
n = 3  # Number of individuals
m = 10 # Number of loci

A = NULL
B = NULL

for(i in 1:m)
{
  A = cbind(A,sample(x=1:S[i],size=n,replace=T))
  B = cbind(B,sample(x=1:S[i],size=n,replace=T))
}

# Makes individual 2 and 3 be siblings
ind = sample(c(F,T),size=m,replace=T)
A[2,ind] = A[3,ind]
ind = sample(c(F,T),size=m,replace=T)
B[2,ind] = B[3,ind]

AB = matrix(paste(A,B,sep="/"),ncol=m,byrow=F)
colnames(AB)=1:m
rownames(AB)=paste("Individ",1:n,sep="")
```

C.2 Computation of LOD Scores

The following code shows the computation of the parent-offspring LOD score, LOD_p, the identity LOD score, LOD_{id} and the full siblings LOD score, LOD_{sib}, for individuals 1 and 2. The data was registered as genetic data by using commands from the package gstudio (Dyer, 2012), available in R.

```
#Gentetic data registered
require(gstudio) #The package gstudio contains the function Locus()

aloc1<-(Locus(c(4,8),phased=FALSE))
aloc2<-(Locus(c(15,6),phased=FALSE))
aloc3<-(Locus(c(8,3),phased=FALSE))
aloc4<-(Locus(c(3,13),phased=FALSE))
aloc5<-(Locus(c(7,10),phased=FALSE))
aloc6<-(Locus(c(8,1),phased=FALSE))
aloc7<-(Locus(c(2,5),phased=FALSE))
aloc8<-(Locus(c(3,9),phased=FALSE))
aloc9<-(Locus(c(6,13),phased=FALSE))
aloc10<-(Locus(c(12,10),phased=FALSE))

bloc1<-(Locus(c(8,2),phased=FALSE))
bloc2<-(Locus(c(12,15),phased=FALSE))
bloc3<-(Locus(c(17,11),phased=FALSE))
bloc4<-(Locus(c(4,6),phased=FALSE))
bloc5<-(Locus(c(8,3),phased=FALSE))
bloc6<-(Locus(c(8,8),phased=FALSE))
bloc7<-(Locus(c(11,1),phased=FALSE))
bloc8<-(Locus(c(10,9),phased=FALSE))
bloc9<-(Locus(c(10,2),phased=FALSE))
bloc10<-(Locus(c(13,16),phased=FALSE))

#The genetic profiles of Individual 1 and 2:
Ind_1<-c(aloc1,aloc2,aloc3,aloc4,aloc5,aloc6,aloc7,aloc8,aloc9,aloc10)
Ind_2<-c(bloc1,bloc2,bloc3,bloc4,bloc5,bloc6,bloc7,bloc8,bloc9,bloc10)

#Allele frequencies
freq1<-0.11111111
freq2<-0.05555556
freq3<-0.05000000
freq4<-0.07142857
freq5<-0.10000000
freq6<-0.08333333
freq7<-0.08333333
freq8<-0.09090910
freq9<-0.05882353
freq10<-0.0625000
```

After the genetic data has been registered then parent-offspring LOD score is computed:

```
#H1: Ind_1 and Ind_2 are a parent and his/her offspring

LRP_1<-0 #Parent-offspring likelihood ratio for locus 1
LRP_1= (0.25*((if(aloc1[1]==bloc1[1]){1}else{0})
+(if(aloc1[2]==bloc1[1]){1}else{0})
+(if(aloc1[1]==bloc1[2]){1}else{0})
+(if(aloc1[2]==bloc1[2]){1}else{0})))/freq1

LRP_2<-0 #Parent-offspring likelihood ratio for locus 2
LRP_2=(0.25*((if(aloc2[1]==bloc2[1]){1}else{0})
+(if(aloc2[2]==bloc2[1]){1}else{0})
+(if(aloc2[1]==bloc2[2]){1}else{0})
+(if(aloc2[2]==bloc2[2]){1}else{0})))/freq2

LRP_3<-0 #Parent-offspring likelihood ratio for locus 3
LRP_3=0.25*(((if(aloc3[1]==bloc3[1]){1}else{0})
+(if(aloc3[2]==bloc3[1]){1}else{0})
+(if(aloc3[1]==bloc3[2]){1}else{0})
+(if(aloc3[2]==bloc3[2]){1}else{0})))/freq3

LRP_4<-0 #Parent-offspring likelihood ratio for locus 4
LRP_4=0.25*(((if(aloc4[1]==bloc4[1]){1}else{0})
+(if(aloc4[2]==bloc4[1]){1}else{0})
+(if(aloc4[1]==bloc4[2]){1}else{0})
+(if(aloc4[2]==bloc4[2]){1}else{0})))/freq4

LRP_5<-0 #Parent-offspring likelihood ratio for locus 5
LRP_5=0.25*(((if(aloc5[1]==bloc5[1]){1}else{0})
+(if(aloc5[2]==bloc5[1]){1}else{0})
+(if(aloc5[1]==bloc5[2]){1}else{0})
+(if(aloc5[2]==bloc5[2]){1}else{0})))/freq5

LRP_6<-0 #Parent-offspring likelihood ratio for locus 6
LRP_6=0.25*(((if(aloc6[1]==bloc6[1]){1}else{0})
+(if(aloc6[2]==bloc6[1]){1}else{0})
+(if(aloc6[1]==bloc6[2]){1}else{0})
+(if(aloc6[2]==bloc6[2]){1}else{0})))/freq6

LRP_7<-0 #Parent-offspring likelihood ratio for locus 7
LRP_7=0.25*(((if(aloc7[1]==bloc7[1]){1}else{0})
+(if(aloc7[2]==bloc7[1]){1}else{0})
+(if(aloc7[1]==bloc7[2]){1}else{0})
+(if(aloc7[2]==bloc7[2]){1}else{0})))/freq7

LRP_8<-0 #Parent-offspring likelihood ratio for locus 8
```

```
LRP_8=0.25*(((if(aloc8[1]==bloc8[1]){1}else{0})
+(if(aloc8[2]==bloc8[1]){1}else{0})
+(if(aloc8[1]==bloc8[2]){1}else{0})
+(if(aloc8[2]==bloc8[2]){1}else{0})))/freq8

LRP_9<-0 #Parent-offspring likelihood ratio for locus 9
LRP_9=0.25*(((if(aloc9[1]==bloc9[1]){1}else{0})
+(if(aloc9[2]==bloc9[1]){1}else{0})
+(if(aloc9[1]==bloc9[2]){1}else{0})
+(if(aloc9[2]==bloc9[2]){1}else{0})))/freq9

LRP_10<-0 #Parent-offspring likelihood ratio for locus 10
LRP_10=0.25*(((if(aloc10[1]==bloc10[1]){1}else{0})
+(if(aloc10[2]==bloc10[1]){1}else{0})
+(if(aloc10[1]==bloc10[2]){1}else{0})
+(if(aloc10[2]==bloc10[2]){1}else{0})))/freq10

LOD_p<-0 #Parent-offspring LOD score computed
LOD_p=log(LRP_1*LRP_2*LRP_3*LRP_4*LRP_5
           *LRP_6*LRP_7*LRP_8*LRP_9*LRP_10,base=10)
```

The identity LOD score is computed in the following way:

```
#H1: Ind_1 and Ind_2 are identical twins.

LRid_1<-0 #Identity likelihood ratio for locus 1
LRid_1= 0.5*(freq1)^(-2)*
(((if(aloc1[1]==bloc1[1]){1}else{0})
*(if(aloc1[2]==bloc1[2]){1}else{0}))
+((if(aloc1[1]==bloc1[2]){1}else{0})
*(if(aloc1[2]==bloc1[1]){1}else{0})))

LRid_2<-0 #Identity likelihood ratio for locus 2
LRid_2=0.5*(freq2)^(-2)*
(((if(aloc2[1]==bloc2[1]){1}else{0})
*(if(aloc2[2]==bloc2[2]){1}else{0}))
+((if(aloc2[1]==bloc2[2]){1}else{0})
*(if(aloc2[2]==bloc2[1]){1}else{0})))

LRid_3<-0 #Identity likelihood ratio for locus 3
LRid_3=0.5*(freq3)^(-2)*
(((if(aloc3[1]==bloc3[1]){1}else{0})
*(if(aloc3[2]==bloc3[2]){1}else{0}))
+((if(aloc3[1]==bloc3[2]){1}else{0})
*(if(aloc3[2]==bloc3[1]){1}else{0})))

LRid_4<-0 #Identity likelihood ratio for locus 4
```

```
LRid_4=0.5*(freq4)^(-2)*
(((if(aloc4[1]==bloc4[1]){1}else{0})
*(if(aloc4[2]==bloc4[2]){1}else{0}))
+((if(aloc4[1]==bloc4[2]){1}else{0})
*(if(aloc4[2]==bloc4[2]){1}else{0})))

LRid_5<-0 #Identity likelihood ratio for locus 5
LRid_5=0.5*(freq5)^(-2)*
(((if(aloc5[1]==bloc5[1]){1}else{0})
*(if(aloc5[2]==bloc5[2]){1}else{0}))
+((if(aloc5[1]==bloc5[2]){1}else{0})
*(if(aloc5[2]==bloc5[1]){1}else{0})))

LRid_6<-0 #Identity likelihood ratio for locus 6
LRid_6=0.5*(freq6)^(-2)*
(((if(aloc6[1]==bloc6[1]){1}else{0})
*(if(aloc6[2]==bloc6[1]){1}else{0}))
+((if(aloc6[1]==bloc6[2]){1}else{0})
*(if(aloc6[2]==bloc6[2]){1}else{0})))

LRid_7<-0 #Identity likelihood ratio for locus 7
LRid_7=0.5*(freq7)^(-2)*
(((if(aloc7[1]==bloc7[1]){1}else{0})
*(if(aloc7[2]==bloc7[2]){1}else{0}))
+((if(aloc7[1]==bloc7[2]){1}else{0})
*(if(aloc7[2]==bloc7[1]){1}else{0})))

LRid_8<-0 #Identity likelihood ratio for locus 8
LRid_8=0.5*(freq8)^(-2)*
(((if(aloc8[1]==bloc8[1]){1}else{0})
*(if(aloc8[2]==bloc8[2]){1}else{0}))
+((if(aloc8[1]==bloc8[2]){1}else{0})
*(if(aloc8[2]==bloc8[1]){1}else{0})))

LRid_9<-0 #Identity likelihood ratio for locus 9
LRid_9=0.5*(freq9)^(-2)*
(((if(aloc9[1]==bloc9[1]){1}else{0})
*(if(aloc9[2]==bloc9[2]){1}else{0}))
+(if(aloc9[1]==bloc9[2]){1}else{0})
*(if(aloc9[2]==bloc9[1]){1}else{0}))

LRid_10<-0 #Identity likelihood ratio for locus 10
LRid_10=0.5*(freq10)^(-2)*
(((if(aloc10[1]==bloc10[1]){1}else{0})
*(if(aloc10[2]==bloc10[2]){1}else{0}))
+((if(aloc10[1]==bloc10[2]){1}else{0})
*(if(aloc10[2]==bloc10[1]){1}else{0})))
```

```
LOD_id<-0 #Identity LOD score computed
LOD_id=log(LRid_1*LRid_2*LRid_3*LRid_4*LRid_5
           *LRid_6*LRid_7*LRid_8*LRid_9*LRid_10,base=10)
```

When LOD_p and LOD_{id} have been computed then computing the LOD scores for other relatedness hypothesis, $H1$, is a simple task by using formula 4.6:

$$LR_{H_1} = k_0(H_1) + k_1(H_1) \cdot LR_p + k_2(H_1) \cdot LR_{id}$$

k_0, k_1 and k_2 are kinship coefficients given the relatedness that is being tested, (see table: 3.1):

```
#H1: Ind_1 and Ind_2 are siblings

LRsib_1<-0 #Siblings likelihood ratio for locus 1
LRsib_1=1/4+1/2*(LRP_1)+1/4*(LRid_1)

LRsib_2<-0 #Siblings likelihood ratio for locus 2
LRsib_2=1/4+1/2*(LRP_2)+1/4*(LRid_2)

LRsib_3<-0 #Siblings likelihood ratio for locus 3
LRsib_3=1/4+1/2*(LRP_3)+1/4*(LRid_3)

LRsib_4<-0 #Siblings likelihood ratio for locus 4
LRsib_4=1/4+1/2*(LRP_4)+1/4*(LRid_4)

LRsib_5<-0 #Siblings likelihood ratio for locus 5
LRsib_5=1/4+1/2*(LRP_5)+1/4*(LRid_5)

LRsib_6<-0 #Siblings likelihood ratio for locus 6
LRsib_6=1/4+1/2*(LRP_6)+1/4*(LRid_6)

LRsib_7<-0 #Siblings likelihood ratio for locus 7
LRsib_7=1/4+1/2*(LRP_7)+1/4*(LRid_7)

LRsib_8<-0 #Siblings likelihood ratio for locus 8
LRsib_8=1/4+1/2*(LRP_8)+1/4*(LRid_8)

LRsib_9<-0 #Siblings likelihood ratio for locus 9
LRsib_9=1/4+1/2*(LRP_9)+1/4*(LRid_9)

LRsib_10<-0 #Siblings likelihood ratio for locus 10
LRsib_10=1/4+1/2*(LRP_10)+1/4*(LRid_10)

LOD_sib<-0 #Siblings LOD score computed
LOD_sib=log(LRsib_1*LRsib_2*LRsib_3*LRsib_4*LRsib_5
            *LRsib_6*LRsib_7*LRsib_8*LRsib_9*LRsib_10,base=10)
```

The half-sibling LOD score and the first cousin LOD score are computed in the same way as the full sibling LOD score, only with different kinship coefficients:

```
LRh-sib_1<-0 #Half-siblings likelihood ratio for locus 1
LRh-sib_1=1/2+1/2*(LRP_1)+0*(LRid_1)

LRcous_1<-0 #First cousins likelihood ratio for locus 1
LRcous_1=3/4+1/4*(LRP_1)+0*(LRid_1)
```

C.3 Estimation of p-Values

The p-values are estimated by simulating 100 unrelated individuals with the same code as can be found in C.1 with

```
n = 100
```

and without the command that makes individual 2 and 3 be siblings. The genetic data is registered a little bit different here. Genotype (A, B) at locus s for individual i is denoted by:

```
A[i,s] #Allele A at locus s for individual i
B[i,s] #Allele B at locus s for individual i
```

When the individuals have been simulated then their pairwise LOD scores are computed in a matrix. The sibling LOD matrix for the simulated individuals is computed in the following way:

```
simLOD_sib<-NULL

simLOD_sib=matrix(ncol=100,nrow=100)

for(i in 1:ncol(simLOD_sib))
{
for(j in 1:nrow(simLOD_sib))
{
simLOD_sib[i,j]=
log(((0.25+0.5*(((if(A[i,1]==A[j,1]){1}else{0})+(if(A[i,1]==B[j,1]){1}else{0})
+(if(B[i,1]==A[j,1]){1}else{0})+(if(B[i,1]==B[j,1]){1}else{0}))
/(4*freq[1]))
+(0.25*(((if(A[i,1]==A[j,1]){1}else{0})*(if(B[i,1]==B[j,1]){1}else{0}))
+((if(B[i,1]==A[j,1]){1}else{0})+(if(A[i,1]==B[j,1]){1}else{0})))
/(2*freq[1]*freq[1])))

*(0.25+0.5*(((if(A[i,2]==A[j,2]){1}else{0})+(if(A[i,2]==B[j,2]){1}else{0})
+(if(B[i,2]==A[j,2]){1}else{0})+(if(B[i,2]==B[j,2]){1}else{0}))
/(4*freq[2]))
+(0.25*(((if(A[i,2]==A[j,2]){1}else{0})*(if(B[i,2]==B[j,2]){1}else{0}))
+((if(B[i,2]==A[j,2]){1}else{0})+(if(A[i,2]==B[j,2]){1}else{0})))
/(2*freq[2]*freq[2])))
```

```
*(0.25+0.5*(((if(A[i,3]==A[j,3]){1}else{0})+(if(A[i,3]==B[j,3]){1}else{0})
+(if(B[i,3]==A[j,3]){1}else{0})+(if(B[i,3]==B[j,3]){1}else{0}))
/(4*freq[3]))
+(0.25*(((if(A[i,3]==A[j,3]){1}else{0})*(if(B[i,3]==B[j,3]){1}else{0}))
+((if(B[i,3]==A[j,3]){1}else{0})+(if(A[i,3]==B[j,3]){1}else{0})))
/(2*freq[3]*freq[3])))

*(0.25+0.5*(((if(A[i,4]==A[j,4]){1}else{0})+(if(A[i,4]==B[j,4]){1}else{0})
+(if(B[i,4]==A[j,4]){1}else{0})+(if(B[i,4]==B[j,4]){1}else{0}))
/(4*freq[4]))
+(0.25*(((if(A[i,4]==A[j,4]){1}else{0})*(if(B[i,4]==B[j,4]){1}else{0}))
+((if(B[i,4]==A[j,4]){1}else{0})+(if(A[i,4]==B[j,4]){1}else{0})))
/(2*freq[4]*freq[4])))

*(0.25+0.5*(((if(A[i,5]==A[j,5]){1}else{0})+(if(A[i,5]==B[j,5]){1}else{0})
+(if(B[i,5]==A[j,5]){1}else{0})+(if(B[i,5]==B[j,5]){1}else{0}))
/(4*freq[5]))
+(0.25*(((if(A[i,5]==A[j,5]){1}else{0})*(if(B[i,5]==B[j,5]){1}else{0}))
+((if(B[i,5]==A[j,5]){1}else{0})+(if(A[i,5]==B[j,5]){1}else{0})))
/(2*freq[5]*freq[5])))

*(0.25+0.5*(((if(A[i,6]==A[j,6]){1}else{0})+(if(A[i,6]==B[j,6]){1}else{0})
+(if(B[i,6]==A[j,6]){1}else{0})+(if(B[i,6]==B[j,6]){1}else{0}))
/(4*freq[6]))
+(0.25*(((if(A[i,6]==A[j,6]){1}else{0})*(if(B[i,6]==B[j,6]){1}else{0}))
+((if(B[i,6]==A[j,6]){1}else{0})+(if(A[i,6]==B[j,6]){1}else{0})))
/(2*freq[6]*freq[6])))

*(0.25+0.5*(((if(A[i,7]==A[j,7]){1}else{0})+(if(A[i,7]==B[j,7]){1}else{0})
+(if(B[i,7]==A[j,7]){1}else{0})+(if(B[i,7]==B[j,7]){1}else{0}))
/(4*freq[7]))
+(0.25*(((if(A[i,7]==A[j,7]){1}else{0})*(if(B[i,7]==B[j,7]){1}else{0}))
+((if(B[i,7]==A[j,7]){1}else{0})+(if(A[i,7]==B[j,7]){1}else{0})))
/(2*freq[7]*freq[7])))

*(0.25+0.5*(((if(A[i,8]==A[j,8]){1}else{0})+(if(A[i,8]==B[j,8]){1}else{0})
+(if(B[i,8]==A[j,8]){1}else{0})+(if(B[i,8]==B[j,8]){1}else{0}))
/(4*freq[8]))
+(0.25*(((if(A[i,8]==A[j,8]){1}else{0})*(if(B[i,8]==B[j,8]){1}else{0}))
+((if(B[i,8]==A[j,8]){1}else{0})+(if(A[i,8]==B[j,8]){1}else{0})))
/(2*freq[8]*freq[8])))

*(0.25+0.5*(((if(A[i,9]==A[j,9]){1}else{0})+(if(A[i,9]==B[j,9]){1}else{0})
+(if(B[i,9]==A[j,9]){1}else{0})+(if(B[i,9]==B[j,9]){1}else{0}))
/(4*freq[9]))
+(0.25*(((if(A[i,9]==A[j,9]){1}else{0})*(if(B[i,9]==B[j,9]){1}else{0}))
+((if(B[i,9]==A[j,9]){1}else{0})+(if(A[i,9]==B[j,9]){1}else{0})))
```

```
/(2*freq[9]*freq[9])))

*(0.25+0.5*(((if(A[i,10]==A[j,10]){1}else{0})+(if(A[i,10]==B[j,10]){1}else{0})
+(if(B[i,10]==A[j,10]){1}else{0})+(if(B[i,10]==B[j,10]){1}else{0}))/(4*freq[10]))
+(0.25*(((if(A[i,10]==A[j,10]){1}else{0})*(if(B[i,10]==B[j,10]){1}else{0}))
+((if(B[i,10]==A[j,10]){1}else{0})+(if(A[i,10]==B[j,10]){1}else{0})))
/(2*freq[10]*freq[10]))))),base=10)
}
}
```

LOD score matrices for other relatedness hypothesis are obtained in the same way just with different kinship coefficients. The matrix contains each LOD score twice, that is $LOD(i,j)$ and $LOD(j,i)$ that are equal, and all LOD scores on the diagonal, $LOD(i,i)$, are just computed LOD scores for the hypothesis that a individual is related to him/her self. Vector containing the relevant LOD scores can be attained by:

```
simLOD_hsib_vector<-simLOD_hsib[upper.tri(simLOD_hsib)]
```

The p-value for the sibling LOD score for individual 1 and 2 is computed by comparing the simulated LOD scores with their observed LOD score:

```
PP<-numeric(4950)
for(j in 1:length(PP)){
PP[j]=if(simLOD_sib_vec[j]<LOD_sib){0}else{1}}

p_value=sum(PP)/4950
```

Appendix D

R Codes for the Fin Whale Analysis

This section concludes the codes that were used in the analysis of the fin whale registry. All computations were done by using the program R, version 2.14.1, (R Development Core Team, 2011).

D.1 Registration of the Genetic Data

The fin whale data is arranged in a matrix, with 266 rows and 31 columns. The first column contains the names of the fin whales while the first row contains the names of the loci. Each row contains genetic information about one fin whale, and each column contains information about one allele at a certain locus. The genotype (A, B) for individual i at locus s would be denoted with:

```
data[i,2s] #Allele A at locus s for individual i
data[i,2s+1] #Allele B at locus s for individual i
```

D.2 Estimation of Population Allele Frequencies

The allele frequencies are estimated from the dataset, excluding the 22 foetuses. The columns have been named after their locus name, (EV001, EV037, GATA028, GATA053, GATA098, GATA417, GT011, GT023, GT195, GT211, GT271, GT310, GT575, TAA023 and GGAA520), but information at locus 1 is denoted with:

```
EV1A=dat[,2]
EV1B=dat[,3]
```

The frequencies are estimated by using the function:

```
Frequencies()
```

from the R package gstudio (Dyer, 2012), but that requires that the data is registered as genetic profiles by using the function:

```
Locus()
```

The estimated allele frequencies are simply the proportion of how many times a certain allele types appears in the sample.

```
require(gstudio)

EV1<-list(NULL)
for(i in 1:length(EV1A))
{
EV1[i]=Locus(c(EV1A[i],EV1B[i]))
}
freqs_1<-Frequencies(c(EV1))

EV37<-list(NULL)
for(i in 1:length(EV37A))
{
EV37[i]=Locus(c(EV37A[i],EV37B[i]))
}
freqs_2<-Frequencies(c(EV37))

GT011<-list(NULL)
for(i in 1:length(GT011A))
{
GT011[i]=Locus(c(GT011A[i],GT011B[i]))
}
freqs_3<-Frequencies(c(GT011))

GT023<-list(NULL)
for(i in 1:length(GT023A))
{
GT023[i]=Locus(c(GT023A[i],GT023B[i]))
}
freqs_4<-Frequencies(c(GT023))


GT195<-list(NULL)
for(i in 1:length(GT195A))
{
GT195[i]=Locus(c(GT195A[i],GT195B[i]))
}
freqs_5<-Frequencies(c(GT195))
GT211<-list(NULL)

for(i in 1:length(GT211A))
{
GT211[i]=Locus(c(GT211A[i],GT211B[i]))
}
freqs_6<-Frequencies(c(GT211))

GT271<-list(NULL)
for(i in 1:length(GT271A))
```

```
{
GT271[i]=Locus(c(GT271A[i],GT271B[i]))
}
freqs_7<-Frequencies(c(GT271))

GT310<-list(NULL)
for(i in 1:length(GT310A))
{
GT310[i]=Locus(c(GT310A[i],GT310B[i]))
}
freqs_8<-Frequencies(c(GT310))

GT575<-list(NULL)
for(i in 1:length(GT575A))
{
GT575[i]=Locus(c(GT575A[i],GT575B[i]))
}
freqs_9<-Frequencies(c(GT575))

GATA028<-list(NULL)
for(i in 1:length(GATA028A))
{
GATA028[i]=Locus(c(GATA028A[i],GATA028B[i]))
}
freqs_10<-Frequencies(c(GATA028))

GATA053<-list(NULL)
for(i in 1:length(GATA053A))
{
GATA053[i]=Locus(c(GATA053A[i],GATA053B[i]))
}
freqs_11<-Frequencies(c(GATA053))

GATA098<-list(NULL)
for(i in 1:length(GATA098A))
{
GATA098[i]=Locus(c(GATA098A[i],GATA098B[i]))
}
freqs_12<-Frequencies(c(GATA098))

GATA417<-list(NULL)
for(i in 1:length(GATA417A))
{
GATA417[i]=Locus(c(GATA417A[i],GATA417B[i]))
}
freqs_13<-Frequencies(c(GATA417))
```

```
GTAA520<-list(NULL)
for(i in 1:length(GTAA520A))
{
GTAA520[i]=Locus(c(GTAA520A[i],GTAA520B[i]))
}
freqs_14<-Frequencies(c(GTAA520))

TAA023<-list(NULL)
for(i in 1:length(TAA023A))
{
TAA023[i]=Locus(c(TAA023A[i],TAA023B[i]))
}
freqs_15<-Frequencies(c(TAA023))
```

D.3 LOD Scores

The following code shows the computation of the LOD matrix for half-siblings hypothesis in *R*. Each frequency matrix has two columns, the first column contains the allele name while the second column contains the corresponding estimated frequency for that allele. The value of the estimated population allele frequency for allele A at locus s for individual i is obtained by the command:

```
freqs[match(data[i,2s],freqs[,1]),2]
```

The half-sibling LOD score matrix is computed in the following way:

```
LOD<-NULL

LOD=matrix(ncol=265,nrow=265)

for(i in 1:ncol(LOD))
{
for(j in 1:nrow(LOD))
{
(LOD[i,j]=
log(
(((0.5)+(0.5*0.25)
*((((if(data[i,2]==data[j,2]){1}else{0})+(if(data[i,2]==data[j,3]){1}else{0}))
/(freq1[match(data[i,2],freq1[,1]),2]))

+(((if(data[i,3]==data[j,2]){1}else{0})+(if(data[i,3]==data[j,3]){1}else{0}))
/(freq1[match(data[i,3],freq1[,1]),2])))))

*((0.5)+(0.5*0.25)
*((((if(data[i,4]==data[j,4]){1}else{0})+(if(data[i,4]==data[j,5]){1}else{0}))
/(freq2[match(data[i,4],freq2[,1]),2]))

+(((if(data[i,5]==data[j,4]){1}else{0})+(if(data[i,5]==data[j,5]){1}else{0}))
```

```
/(freq2[match(data[i,5],freq2[,1]),2])))

*((0.5)+(0.5*0.25)
*((((if(data[i,6]==data[j,6]){1}else{0})+(if(data[i,6]==data[j,7]){1}else{0}))
/(freq3[match(data[i,6],freq3[,1]),2]))


+(((if(data[i,7]==data[j,6]){1}else{0})+(if(data[i,7]==data[j,7]){1}else{0}))
/(freq3[match(data[i,7],freq3[,1]),2])))

*((0.5)+(0.5*0.25)
*((((if(data[i,8]==data[j,8]){1}else{0})+(if(data[i,8]==data[j,9]){1}else{0}))
/(freq4[match(data[i,8],freq4[,1]),2]))

+(((if(data[i,9]==data[j,8]){1}else{0})+(if(data[i,9]==data[j,9]){1}else{0}))
/(freq4[match(data[i,9],freq4[,1]),2])))

*((0.5)+(0.5*0.25)
*((((if(data[i,10]==data[j,10]){1}else{0})+(if(data[i,10]==data[j,11]){1}else{0}))
/(freq5[match(data[i,10],freq5[,1]),2]))

+(((if(data[i,11]==data[j,10]){1}else{0})+(if(data[i,11]==data[j,11]){1}else{0}))
/(freq5[match(data[i,11],freq5[,1]),2])))

*((0.5)+(0.5*0.25)
*((((if(data[i,12]==data[j,12]){1}else{0})+(if(data[i,12]==data[j,13]){1}else{0}))
/(freq6[match(data[i,12],freq6[,1]),2]))

+(((if(data[i,13]==data[j,12]){1}else{0})+(if(data[i,13]==data[j,13]){1}else{0}))
/(freq6[match(data[i,13],freq6[,1]),2])))

*((0.5)+(0.5*0.25)
*((((if(data[i,14]==data[j,14]){1}else{0})+(if(data[i,14]==data[j,15]){1}else{0}))
/(freq7[match(data[i,14],freq7[,1]),2]))

+(((if(data[i,15]==data[j,14]){1}else{0})+(if(data[i,15]==data[j,15]){1}else{0}))
/(freq7[match(data[i,15],freq7[,1]),2])))

*((0.5)+(0.5*0.25)
*((((if(data[i,16]==data[j,16]){1}else{0})+(if(data[i,16]==data[j,17]){1}else{0}))
/(freq8[match(data[i,16],freq8[,1]),2]))

+(((if(data[i,17]==data[j,16]){1}else{0})+(if(data[i,17]==data[j,17]){1}else{0}))
/(freq8[match(data[i,17],freq8[,1]),2])))

*((0.5)+(0.5*0.25)
*((((if(data[i,18]==data[j,18]){1}else{0})+(if(data[i,18]==data[j,19]){1}else{0}))
```

```
/(freq9[match(data[i,18],freq9[,1]),2]))

+(((if(data[i,19]==data[j,18]){1}else{0})+(if(data[i,19]==data[j,19]){1}else{0}))
/(freq9[match(data[i,19],freq9[,1]),2])))))

*((0.5)+(0.5*0.25)
*((((if(data[i,20]==data[j,20]){1}else{0})+(if(data[i,20]==data[j,21]){1}else{0}))
/(freq10[match(data[i,20],freq10[,1]),2]))

+(((if(data[i,21]==data[j,20]){1}else{0})+(if(data[i,21]==data[j,21]){1}else{0}))
/(freq10[match(data[i,21],freq10[,1]),2])))))

*((0.5)+(0.5*0.25)
*((((if(data[i,22]==data[j,22]){1}else{0})+(if(data[i,22]==data[j,23]){1}else{0}))
/(freq11[match(data[i,22],freq11[,1]),2]))

+(((if(data[i,23]==data[j,22]){1}else{0})+(if(data[i,23]==data[j,23]){1}else{0}))
/(freq11[match(data[i,23],freq11[,1]),2])))))

*((0.5)+(0.5*0.25)
*((((if(data[i,24]==data[j,24]){1}else{0})+(if(data[i,24]==data[j,25]){1}else{0}))
/(freq12[match(data[i,24],freq12[,1]),2]))

+(((if(data[i,25]==data[j,24]){1}else{0})+(if(data[i,25]==data[j,25]){1}else{0}))
/(freq12[match(data[i,25],freq12[,1]),2])))))

*((0.5)+(0.5*0.25)
*((((if(data[i,26]==data[j,26]){1}else{0})+(if(data[i,26]==data[j,27]){1}else{0}))
/(freq13[match(data[i,26],freq13[,1]),2]))

+(((if(data[i,27]==data[j,26]){1}else{0})+(if(data[i,27]==data[j,27]){1}else{0}))
/(freq13[match(data[i,27],freq13[,1]),2])))))

*((0.5)+(0.5*0.25)
*((((if(data[i,28]==data[j,28]){1}else{0})+(if(data[i,28]==data[j,29]){1}else{0}))
/(freq14[match(data[i,28],freq14[,1]),2]))

+(((if(data[i,29]==data[j,28]){1}else{0})+(if(data[i,29]==data[j,29]){1}else{0}))
/(freq14[match(data[i,29],freq14[,1]),2])))))

*((0.5)+(0.5*0.25)
*((((if(data[i,30]==data[j,30]){1}else{0})+(if(data[i,30]==data[j,31]){1}else{0}))
/(freq15[match(data[i,30],freq15[,1]),2]))

+(((if(data[i,31]==data[j,30]){1}else{0})+(if(data[i,31]==data[j,31]){1}else{0}))
/(freq15[match(data[i,31],freq15[,1]),2])))))),base=10))
}
```

```
}
```

The parent-offspring and first cousins LOD matrices are computed in the same way just with different kinship coefficients.

D.4 Simulation of Individuals

The following code was used to simulate 265 unrelated individuals with the allele frequencies that were estimated from the dataset. It is built on the code from C.1 that Hans Julius Skaug wrote. Before the simulation, the matrix that was used to estimated the population allele frequencies, was rearranged by aligning alleles A and B at locus s in the same column. That resulted in a matrix called S, that has 15 columns, one for each locus, and 486 rows, one for each allele that occurs in the dataset (excluding the foetuses). The 265 unrelated individuals were simulated by drawing independently from this matrix with replacement.

```
n = 265  # Number of individuals
m = 15 # Number of loci

A = NULL
B = NULL

for(i in 1:m)
{
  A= cbind(A,sample(S[,i],size=n,replace=T))
  B = cbind(B,sample(S[,i],size=n,replace=T))
}

AB = matrix(paste(A,B,sep="/"),ncol=m,byrow=F)
colnames(AB)=1:m
rownames(AB)=paste("Individ",1:n,sep="")
```

D.5 Computation of p-Values

The LOD scores for the simulated individuals are computed in the same way as the LOD matrices for the real individuals. The matrices contain each LOD score twice, $LOD(i,j)$ and $LOD(j,i)$, and all LOD scores on the diagonal of those matrices, $LOD(i,i)$, are simply computed LOD score for the hypothesis that an individual is related to him/her self. Vectors containing the relevant LOD scores can be attained by:

```
#Vector containing the LOD scores computed from the fin whale dataset
LOD_real<-LOD[upper.tri(LOD)]

#Vector containing LOD scores computed from the simulated dataset
LOD_sim<-simLOD[upper.tri(simLOD)]
```

The p-values for each LOD score are computed in the following way:

```
PP<-NULL
PP=matrix(nrow=length(LOD_sim),ncol=length(LOD_real))
for(j in 1:nrow(PP))
{
for(i in 1:ncol(PP))
{
PP[j,i]=if(LOD_sim[j]<LOD_real[i]){0}else{1}
}
}

p_values<-NULL
p_values=matrix(ncol=length(LOD_real),nrow=2)
p_values[1,]=LOD_real #In order to have the LOD score value with the p-value
for(i in 1:ncol(p_values))
{
p_values[2,i]=sum(PP[,i])/34980
}
```

D.6 Exact Binomial Confidence Interval

The exact confidence intervals are computed here by using the package binom (Dorai-Raj, 2009), available in R.

```
require(binom) #Package available in R

#Computes exact 95% confidence interval for the estimated p-value
#of the parent-offspring LOD score of F09-075 and F10-123
binom.confint(1,2098800,conf.level=0.95,methods="exact")

#Computes exact 95% confidence interval for the estimated p-value
# of the parent-offspring LOD score of F09-125 and F10-119
binom.confint(19,2098800,conf.level=0.95,methods="exact")
```

Appendix E

Estimated Population Allele Frequencies for the Fin Whale Analysis

Table E.1: Allele frequencies at locus 1 to 6

EV001	EV037	GT011	GT023	GT195	GT211
157: 0.3436	187: 0.0782	119: 0.0988	123: 0.0576	161: 0.2058	116: 0.2037
169: 0.2202	193: 0.2901	129: 0.2222	127: 0.1749	169: 0.0905	118: 0.0494
171: 0.0761	189: 0.0021	117: 0.1317	131: 0.0329	171: 0.1934	120: 0.3272
163: 0.1358	191: 0.1255	131: 0.0947	125: 0.1296	179: 0.0288	122: 0.0556
175: 0.0556	179: 0.0021	125: 0.1379	129: 0.3436	173: 0.2654	114: 0.0741
159: 0.0720	181: 0.0247	123: 0.0823	133: 0.1399	175: 0.1276	106: 0.0988
143: 0.0041	183: 0.1214	127: 0.2305	135: 0.0247	177: 0.0576	112: 0.1440
155: 0.0082	197: 0.1770	133: 0.0021	143: 0.0144	167: 0.0226	108: 0.0103
165: 0.0412	199: 0.0658		121: 0.0576	163: 0.0062	126: 0.0021
173: 0.0165	201: 0.0226		109: 0.0144	181: 0.0021	110: 0.0123
161: 0.0062	195: 0.0617		141: 0.0082		124: 0.0226
177: 0.0021	207: 0.0041		97: 0.0021		
145: 0.0041	211: 0.0082				
167: 0.0123	185: 0.0103				
153: 0.0021	205: 0.0021				
	213: 0.0021				
	215: 0.0021				

Table E.2: Allele frequencies at locus 7 to 12

GT271	GT310	GT575	GATA028	GATA053	GATA098
114: 0.2222	110: 0.1955	146: 0.0741	191: 0.0782	246: 0.3848	104: 0.2860
116: 0.4588	112: 0.4588	154: 0.4074	211: 0.0617	266: 0.1070	116: 0.0412
122: 0.0329	114: 0.0329	160: 0.0473	207: 0.1379	260: 0.1276	108: 0.1687
118: 0.1358	122: 0.0309	150: 0.0247	227: 0.2243	270: 0.1111	96: 0.2531
120: 0.0535	126: 0.1584	158: 0.1008	223: 0.0926	262: 0.2016	100: 0.1728
128: 0.0206	124: 0.0700	152: 0.2778	219: 0.0823	258: 0.0226	112: 0.0658
108: 0.0658	130: 0.0021	156: 0.0638	215: 0.0535	250: 0.0309	92: 0.0062
112: 0.0062	120: 0.0309	168: 0.0041	203: 0.0556	274: 0.0103	120: 0.0062
126: 0.0021	106: 0.0185		231: 0.0988	278: 0.0041	
110: 0.0021	118: 0.0021		235: 0.0638		
			199: 0.0350		
			239: 0.0123		
			195: 0.0041		

Table E.3: Allele frequencies at locus 13 to 15

GATA417	GGAA520	TAA023
269: 0.2675	217: 0.0267	95: 0.3889
281: 0.0535	223: 0.1235	101: 0.1667
273: 0.1152	207: 0.1379	86: 0.2510
285: 0.0761	227: 0.0576	92: 0.0350
229: 0.0494	203: 0.0514	98: 0.0802
209: 0.0062	211: 0.1523	104: 0.0761
213: 0.0453	215: 0.1152	89: 0.0021
261: 0.0556	209: 0.0329	
225: 0.0638	201: 0.0309	
237: 0.0123	219: 0.1646	
277: 0.0556	205: 0.0741	
265: 0.1111	231: 0.0226	
217: 0.0391	213: 0.0041	
241: 0.0062	199: 0.0062	
289: 0.0165		
257: 0.0021		
221: 0.0165		
233: 0.0062		
297: 0.0021		