Multiple comparisons of populations based on genetic marker data

Masterarbeit

zur Erlangung des Grades eines M.Sc. der Pflanzenbiotechnologie
der Naturwissenschaftlichen Fakultät
der Leibniz Universität Hannover

vorgelegt von

Name: Schmidt Vorname: Malthe

Geb. am: 18.03.1985 in: Lübeck

Erstprüfer: Prof. Dr. Ludwig Hothorn
Zweitprüfer: Prof. Dr. Bernhard Hau

Hannover, den 06. Dezember 2010

Kurzzusammenfassung

Eine Aufgabe von Populationsgenetikern ist die Untersuchung von Populationen mit dem Ziel, diese in Gruppen zu unterscheiden. Ziel dieser Arbeit ist es, Populationen durch statistisches Testen zu differenzieren. Datengrundlage dazu sind z.B. Analyseergebnisse von genetischen Markern. Diese sind spezifisch für einen Genort und bilden den Allelzustand eines Individuums an diesem Genort unabhängig von Umwelteinflüssen ab. Populationen können verglichen werden, indem zunächst aus einer repräsentativen Stichprobe von Individuen der jeweiligen Population an einem repräsentativen Genort, Allelfrequenzen geschätzt werden. Darauf basierend können die Varianzen innerhalb der Populationen zum Verhältnis zur totalen Varianz als Unterscheidungsmaß genommen und so Populationen unterschieden werden. Diese Idee wird in den nach Weir and Cockerham (1984) beschriebenen F-Statistiken von Wright (1951) in der vorliegenden Arbeit als Maß zur genetischen Differenzierung genutzt. Über das Hardy-Weinberg Gleichgewicht können Genotypfrequenzen für die Population geschätzt und basierend auf der Stichprobe von Individuen zur Berechnung der Teststatistik verwendet werden. Der p-Wert wird zur statistischen Auswertung genutzt. Die dazu nötige Verteilung wird empirisch durch Bootstraps, die über die Populationen stratifiziert sind, erstellt. Die Verteilungen sind, abhängig von der Allelfrequenz, z.T. sehr diskret, was zu einem sehr konservativen oder sehr liberalen p-Wert führt. Das Verfahren kann für den Vergleich zwischen zwei Populationen, den Globalvergleich mehrerer Populationen und den paarweisen Vergleich mehrerer Populationen eingesetzt werden. Bei einem paarweisen Vergleich ist die Implementation einer Multiplizitätsadjustierung nötig. Dazu wird bei dem Bootstrap nur die maximale Bootstrap Teststatistik mit der originalen Teststatistik verglichen. Neben der Kontrolle des globalen Fehlerniveaus wird so auch die Korrelation zwischen den Populationsvergleichen berücksichtigt. Diese Technik stammt von Westfall and Young (1993), wird in dieser Arbeit jedoch angepasst. Anstelle des Resamplings der Residuen in einem linearen Modell wird z.B. die empirische F_{ST}-Verteilung verwendet. Simulationen verschiedener Populationseinstellungen zeigen, dass die Techniken funktionieren und Populationen unterschieden werden können. Die Ergebnisse sind jeodch konservativ und schöpfen das gegebene Fehlerniveau nicht aus. Die Güte ist umso höher, je größer die Populationen sind, je näher die Allelfrequenzen am Rand liegen und je größer die Allelfrequenzunterschiede sind.

Schlagworte: Wright's F_{ST}, empirische F_{ST}-Verteilung, stratifizierter Bootstrap, paarweise Vergleiche, FWER

Abstract

One part of the work of population genetics are analyses of populations with the aim to separate them. The aim of this work is the differentiation of populations by statistical testing. The data basis for the analysis are genetic marker data, which show specific the allelic state of an individual at one locus independent from environmental influences. The populations can be compared by using a representative sample of individuals at a representative part of the genome. Allele frequencies are estimated out of the samples and the proportion of the variance within the populations to the total variance can be used to separate the populations. The implementation of this idea by Weir and Cockerham (1984), which basis on the F-statistics of Wright (1951), is used in this work as a measure of genetic differentiation. With the Hardy-Weinberg Equilibrium the genotype frequencies of the population are estimated from the sampled individuals to calculate the test statistic. The p-value is used for the statistical interpretation. The required distribution is empirically estimated with a over the populations stratified bootstrap. The distributions are, depending on the allele frequencies and sample sizes, very discrete. That leads to a very liberal or very conservative p-value. This method can be used for the comparison of two and multiple populations. Multiple populations are then compared with a global and an all pair comparison. The all pair comparison requires a multiplicity adjustment, so only the maximal test statistic of each bootstrap is compared with the original test statistic. Therefore the family wise error rate is controlled and the correlation between the comparisons is taken into account, too. The usage of the maximal test statistic is adapted from Westfall and Young (1993) but e.g. instead of a normal distribution the empirical F_{ST} distribution is used. The results of the simulations at different settings prove that the methods work and populations can be separated. The results are quite conservative, because the given error rate is hardly used. The greater the population sizes are, the greater the differences of the allele frequencies are and the closer the allele frequencies are to the border of the frequency range, the better is the power to detect a difference between the populations.

Keywords: Wright's F_{ST}, empirical F_{ST} distribution, stratified bootstrap, all pair comparison, FWER

Contents

List of Figures

List of Tables

1 Introduction

1.1 Differentiation of populations

One important topic of populations genetics is the differentiation of populations. The modern genetics deliver distinct information about the allelic state of each individual of a population by genetic markers. With sampled data it is possible to estimate allele frequencies for each sampled population and for all populations together and to use a constitutive statistic to differentiate them. By this statistic, which deals with the relationship of genetic variation within versus among populations, the genetic difference of populations can be quantified. This information can then be used to characterize the genetic structure or gene flow. The greater the difference of the allele frequency between the populations at the analysed locus is, the greater the difference between the populations is. This difference is measured by an appropriate statistic, based e.g. on the allelic state of the sampled individuals. In this work as much genetic variability as possible should be explained by the difference of the genetic information in at least two populations. There are several ways available of comparing populations based on genetic markers, e.g. the comparison of each population at each marker, a conclusion from all markers about each population, a deduction from multiple populations by each marker or a conclusion from all markers about multiple populations. In this work only one marker locus is taken into account by looking at all populations simultaneously. One of several assumptions, which are described later on, is that the genetic marker is on the locus of interest. Based on the marker information populations can be differentiated. The locus should be representative for the genome or the character of interest as well as the individuals for the populations, too. The aim of this work is to compare populations by statistical testing. More precisely to reject H_0, which states that the populations are equal, through the F_{ST} statistic to separate the populations. An illustration of the dispartment of genetic variability is shown in Figure 1. There are the total population level (T), the subdivisions (S) and the individuals (I). Each exemplary individual is considered to have a biallelic locus which is reviewed. This individual can have one of three genotypes, indicated by a white, gray or black circle. The classification of the individuals and their genotypes can be done e.g. with a molecular marker.

After this short introduction a motivating example is presented, a definition of populations and an explanation of molecular markers is shown. In Chapter 2 the assumptions and parameters of interest are mentioned. Chapter 3 to 5 explain the basis for simulating the comparison of two populations, as well as multiple populations at a global comparison and at an all pair comparison. The simulations are presented in Chapter 6. After the behavior of the test statistic is known, it is used on example data sets in Chapter 7. In the end a general discussion in Chapter 8 and at last the appendix, e.g. with the used R-code, are presented.

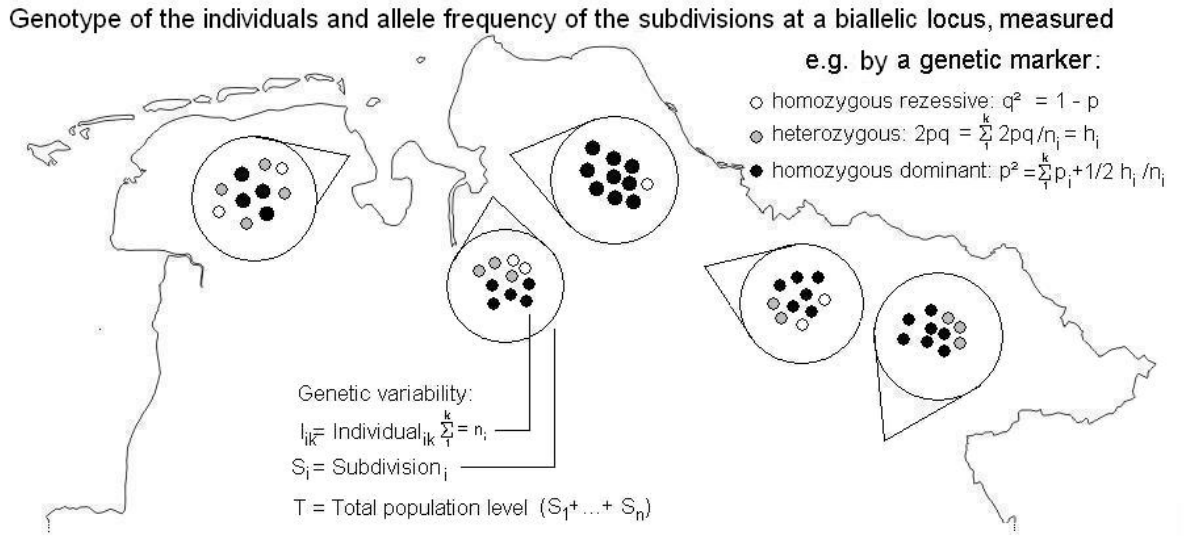

Figure 1: Genetic variability through individuals (I) , subdivisions (S) and the total population level (T) of hypothetical samples.

1.2 Motivating example

An example data set, that could be analysed, comes from a trial done by A.-K. Lühmann, Ph.D. from the Leibniz Universität Hannover. The original data includes 22 populations from 14 places, extracts of which are shown in Table 1. The data derives from a haploid fungus with an unknown reproduction system. It can also be asexual which would ignore the Hardy-Weinberg Equilibrium (HWE) assumption of random mating, which is described in Section 2.1.1. The used markers are originally codominant but were evaluated as a dominant marker by the experimenter. Furthermore, many values are missing. It is of interest if the sampled populations differ from each other. The fungus populations should differ from each other, because of low migration between the areas where the populations are gathered, drift, mutation and selection but also due to other effects that can occur in populations (Balloux and Lugon-Moulin, 2002). These points do not make this data set the first choice for an example data set for this work, but a start to show the structure of data that has to be analysed. It also points out possible problems, which could violate assumptions. The analysis of the motivating example is shown in Section 7.

Table 1: Extract form the data set. It contains information about the number of individuals, the main sampling location and the results of a dominant evaluated genetic marker of 16 loci, coded by 0 or 1.

Individual	Location	Locus1	...	Locus16
1	Bremen	1	...	0
2	Bremen	1	...	0
3	Bremen	1	...	1
4	Bremen	0	...	NA
5	Bremen	1	...	NA
⋮	⋮	⋮		⋮
409	Dortmund	1	...	NA
410	Dortmund	1	...	NA
⋮	⋮	⋮		⋮
550	Dortmund	1	...	1
551	Dortmund	1	...	1
⋮	⋮	⋮		⋮
716	Lauenau	1	...	0

1.3 Populations

There is more than one definition for populations and population structure. A list of 18 definitions out of four backgrounds from ecological, evolutionary and statistical paradigms or through variations of these is presented by Waples and Gaggiotti (2006). The definition of a population changed over the years and is now defined as the totality of individuals of a species in the same area at the same time. Every individual contributes to the common gene pool. The smallest population structure can be called a 'deme' (Balloux and Lugon-Moulin, 2002) or management unit (MU) (Moritz, 1994). The definition of the MU is based on a significant divergence of allele frequency. The alleles can originate from nuclear or mitochondrial loci. Their importance on phylogenetic characteristics depends on the experimenters opinion. Also not crucial alleles can be used. Palsboll et al. (2007) adds that independent populations exist, because of local birth and death rates rather than because of immigration. Rejecting panmixia as the criterion for management units instead of the genetic divergence can lead to misinterpretation in some cases, so it is useful to assign MUs by a threshold level of dispersal. Currently there is no general threshold value or dispersal rate to define populations as separated or not (Waples and Gaggiotti, 2006). There are different methods to get values which can be used to classify populations. One way is to use 'gene flow', the number of migrants per generation which is estimated

by

$$mN_i \tag{1}$$

where m is the probability that an individual is a migrant and N_i are the effective population sizes (Wright, 1951; Palsboll et al., 2007; Balding et al., 2001). Waples and Gaggiotti (2006) described the migration rate (m) as the ecological paradigm and migrants per generation (Nm) as the evolutionary paradigm. In most cases these numbers are unknown and so methods are required to infer mN_i (Leviyang, 2010). These are described in Section 2.2.2 and use the same statistic used to separate the populations.

1.4 Molecular marker

Different types of markers are available. There are phenotype, protein and DNA markers. They have a common condition or appearance that correlates with one or more genes. A phenotype marker can be e.g. blossom colour with the simplifying assumption that it is under control of just one gene. An example could be Figure 2 where at a biallelic locus each genotype would have one colour (incomplete dominant) or two colours (dominant recessive), if a phenotype marker is used. A dominant allele would cover the recessive allele. Phenotype and protein markers are rare and the results are sometimes not distinct e.g. in reference to differentiate between the allelic state (Kumar et al., 2009). Morphology based markers become unexact under environmental influences. Protein markers are unable to detect low levels of variation, because they only screen the protein coding regions of the genome but they allow a separation between the homozygous genotype and the heterozygous genotypes (Laurentin, 2009). Molecular markers are DNA-based and overcome the previously listed disadvantages. DNA markers are distinct and almost unlimitedly available (Laurentin, 2009). Each molecular marker system uses different regions in the genome. Mutation rates are not evenly distributed, so that comparisons with different molecular markers can have varying results depending on the marker regions (Holsinger and Weir, 2009). Depending on the evolutionary distance a molecular marker with the corresponding mutation rate can be used. The use and number of genetic markers has increased fundamentally since their particular invention (Palsboll et al., 2007; Kumar et al., 2009). Advantages and disadvantages of many molecular markers are also reported in Kumar et al. (2009). They also describe an 'ideal' molecular marker. It is highly polymorphic so that differences between the populations can be shown (1), is codominant to avoid calculations for allele frequencies under strong assumptions (2), has an even distribution over the genome (unless only a single gene is of interest) (3) and is selective neutral so that e.g. no environmental methylation can effect the marker (4). Also the marker detection is easy, fast and cheap (5) and trial replications deliver the same result a marker with high reproducibility would do (6). Molecular markers can be separated in many ways. One common possibility is the separation in non PCR based

markers (e.g. RFLP) and PCR based markers (e.g. RAPD, microsatellite, etc.) the other commonly used classification is the separation into dominant and codominant markers.

1.5 Dominant and codominant markers

Molecular markers can be classified into dominant and codominant markers. Dominant markers can only show one homozygote genotype, in most cases the recessive (aa) one. The homozygote dominant (AA) has the same result as a heterozygote genotype (Aa) individual. The at the same time shown allele frequencies must be calculated with the HWE, so the frequency of the heterozygous genotype cannot be determined directly which leads to an approximation. This marker system is much less informative and a divergence of the population to the HWE caused by drift can not be shown. If the marker is not dominant it is compellent codominant.
A codominant marker delivers distinct results for each genotype and the allele frequencies must not be calculated under the HWE and its assumptions. Heterozygote individuals can be distinguished from homozygote individuals. This marker type delivers much more information than the dominant marker, so deviations of populations from the HWE can be detected.
One example of a nearly 'ideal' marker are microsatellite markers. They are codominant and the distribution depends on the marker used. Microsatellite markers are selective neutral and offer a good detection and a high reproducibility. Disadvantages are the unknown mutation rate for a specific locus and the fact that the mutation rate differs between the microsatellite markers. The repeat types and base composition of the repeat can differ between microsatellite type and taxonomic group (Bachtrog et al., 2000). Other influencing factors are the position on the chromosome, length of the allele and whether or not they have flanking sequences (Balloux and Lugon-Moulin, 2002). Their mutation rate is approximately at 10^{-3} which is quite high in comparison to other molecular markers (Weber and Wong, 1993; Jarne and Lagoda, 1996). The drawback is their pricy production, because of the required effort to produce them.

2 Model

2.1 Assumptions

2.1.1 Hardy-Weinberg principle

The mathematician G. H. Hardy and the physicist W. Weinberg developed a way to describe allele frequencies in a population independently from each other, in 1908. In fact the law of equilibrium in absence of selection was already pronounced by W.E. Castle in 1903. He verified the 1:2:1 distribution by U. Yule but eliminated some mistakes

(Aichinger and Grimm, 2008). The basic of the Hardy-Weinberg principle are Mendel's inheritance laws. The standard example to describe the Hardy-Weinberg equilibrium (HWE) is a biallelic locus where p stands for the frequency of the dominant allele A and q for the frequency of the recessive allele a.

$$(p+q)*(p+q) = p^2 + 2pq + q^2 = 1 \tag{2}$$

$$p + q = 1 \tag{3}$$

The dominant allele A and the recessive allele a have the frequency p and q, totaled they add up to 1. In this case of a dominant recessive gene, there are three possible genotypes: AA, Aa and aa with the frequency p^2, 2pq and q^2. The equilibrium, where the allele frequencies of the offspring are the same as in the parental generation, has some strong requirements. There is no type of mutation to the genome of each individual, no migration into or out of the population and no selection which would benefit or handicap any allele. Genetic drift or the bottle neck effect is excluded at the HWE, because a bulk of the population would be eliminated. As a result the remaining individuals would represent different allele frequencies than the gene pool before, because of the reduced population size. In small populations inbreeding occurs more often resulting in an increasing amount of homozygote individuals. The last assumption of the HWE is random mating, so each individual has the same chance to mate with a random individual. The exclusion of non random mating, drift and inbreeding requires a large population. If these assumptions are fulfilled, an ideal population is the result and the HWE will be obtained after one generation of random mating, because the allele frequencies are constant. It is for example possible to calculate the frequency of the heterozygous genotype for each allele, if the frequency of the homozygous individuals is known. This relation is shown in Figure 2 and is a popular scheme which is shown in many publications, as e.g. in Aichinger and Grimm (2008).

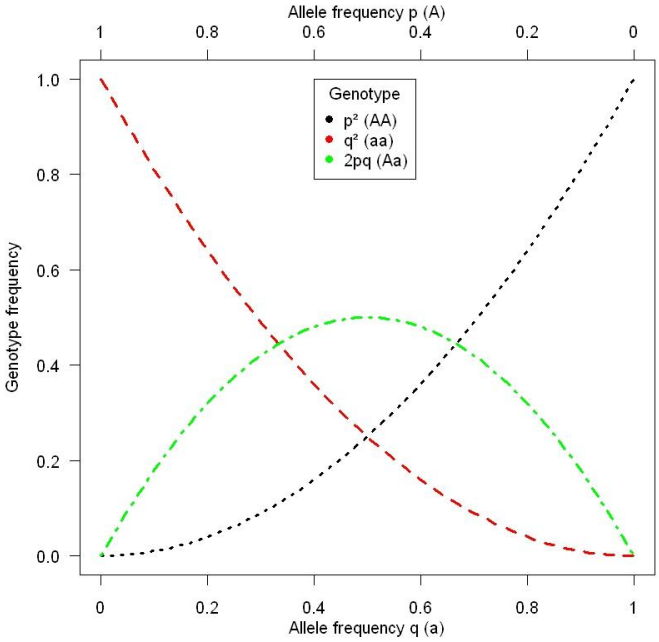

Figure 2: Genotype frequencies as a function of allele frequency in Hardy-Weinberg equilibrium, at a biallelic locus.

2.1.2 Falconer's additive model

The statistic, that will be used later on, includes sums of different terms of variance. The principle of adding terms of variance was adapted for breeding trials by Falconer (1970) from Fisher's additive model (Fisher, 1930). In the model it is assumed that the sum of the variance within populations and between populations is the total variance. It bases on the model of the additive genotype or additive effect of genes. If one locus is considered the genotype value is an addition of the breeding value and the dominance deviation. The breeding values for the genotypes are shown in Table 2. The parameter β denotes the slope of the regression line of the number of A in a genotype to the breeding value.

Table 2: Breeding values for the genotypes at a biallelic locus under HWE in the Falconer model.

Genotype	Breeding value
AA	$2\beta_1 = 2\mathrm{p}\beta$
Aa	$\beta_1+\beta_2 = (\mathrm{p\text{-}q})\beta$
aa	$2\beta_2 = \text{-}2\mathrm{q}\beta$

At HWE the breeding value is null otherwise the population is not in HWE:

$$2q^2p\beta \; + \; 2pq(p-q)\beta \; - \; 2p^2q\beta \; = \; 2pq\beta(p+q-p-q) \; = \; 0$$

The breeding value is also called the additive genotype which stands for the additive effect of genes. At a single locus G the model is reduced by epistatic effects and can be written as

$$G = A$$

where A is the sum of breeding values. Analogue to G, the F statistics F_{ST}, F_{IT} and F_{IS} are assembled by the ratio of the sum of variance terms under the assumptions of the Falconer Model.

2.1.3 Mutation models

Usually, at more complicated settings, different mutation models can be assumed. The simple setting here with a single biallelic locus with allele A and a, HWE and a single generation simplifies the mutation model. A mutation can change A to a or a to A.

2.2 Parameter of interest

2.2.1 Allele frequency

For the comparison of populations, applicable data are allele frequencies e.g. by a genetic marker this data can be generated or observed for each population. With a classification of each individual to a population it is possible to sum the information of the allelic state, to get an population estimate. The allele frequency for the whole population can be obtained by the genotypes of the individuals of each subpopulation. With this information one important value, the frequency of the dominant allele p, is known if the sums of the homozygote dominant genotype AA is added to one half of the heterozygote Aa. The frequency q of the recessive allele is easy to obtain with Equation 3. The proportion of the heterozygous individuals can be obtained under the assumption of HWE, too. Four important values are available by counting and summarizing marker data. The first is the state of the genetic marker of individual ind_k with $k = 1, \ldots, n$. The calculations of the other three, the allele frequencies, genotype frequencies and population sizes are based on the observed individuals. Let I be the number of the compared populations and index i has the range $1, \ldots, I$. Information about the frequency of the dominant allele p_i, the proportion of the heterozygous individuals of the population h_i and the size n_i of the population, can be obtained by the individuals of the populations. The used R code to estimate $\hat{p_i}$, $\hat{h_i}$ is shown in Section A.1.2.

2.2.2 Wright's F-statistics

The aim to differentiate populations, can be accomplished with several statistics. Population differentiation parameters are the method of choice, because a value of differentiation

is presented which shows the correlation or relation of the populations. One of these parameters are Wright's F-statistics (Wright, 1951), which were invented by Sewall Wright in the 1920s to separate cattle breeds. Although the idea is similar to a simple one way ANOVA (analysis of variance), simplified adding variances to a total variance, it is not the same as the ANOVA F statistics, even though they are both called F statistics. They were introduced to analyse the genetic structure of diploid populations, by the departure of genotype frequencies from HWE, more precisely the divergence of heterozygosity (Holsinger and Weir, 2009). Wright's F-statistics are put together through the F coefficients. The F coefficients T, S and I allocate the genetic variability to the total population level (T), to the subdivisions (S) and to the individuals (I) (Hedrick, 2000). An illustration of T, S and I is shown in Figure 1. I and S together from the inbreeding coefficient F_{IS} or simply f. It is a measure of the size of genetic inbreeding within subpopulations and can range from –1, if all individuals are heterozygous, to +1, if there are no observed heterozygous individuals. The overall fixation index is F_{IT} or F. It describes the correlation of alleles between individuals and the population, more precisely the mean deviance of genotype frequencies from the HWE at the total population level (Holsinger and Weir, 2009). The last of the three F statistics is F_{ST} or θ. F_{ST} is the 'most commonly used measure of genetic divergence among populations' (Palsboll et al., 2007; Holsinger and Weir, 2009). It compares the variance of the allele frequencies within the populations with the total population variance of allele frequencies and thus separates populations. It can also be defined by allele correlations of subpopulations relative to the total population (Holsinger and Weir, 2009) or as the extent of divergence among populations relative diversity within the species (Charlesworth, 1998). It can take values between null, if the populations are identical, and one if the subpopulations have completely different fixed alleles or in other words are completely separated (Myles et al., 2008). Originally it was defined by Wright to take positive values only (Hedrick, 2000). It gives information about the correlation between two alleles, chosen at random within subpopulations, relative to alleles sampled at random from the total population (Wright, 1951, 1965). Therefore, it measures inbreeding due to the correlation among alleles, because they are found in the same subpopulation or more precisely the heterozygote deficit relative to its expectation under HWE is measured. Even if a codominant marker is used, the HWE is still assumed by F_{ST}. The relationship of the F statistics is (Balding et al., 2001):

$$(1 - F_{IT}) = (1 - F_{IS})(1 - F_{ST})$$

Each of the F-statistics has a certain area of application. There are different ways to estimate a statistic from data like maximum-likelihood estimates, Bayesian estimates (Pearse and Crandall, 2004) and method-of-moments estimates as F_{ST} (Weir and Cockerham, 1984). It needs parameters from the populations, as the frequency of the dominant allele

p_i and the frequency of the heterozygous genotype h_i and at last the size n_i. The implementation of Wright's F-statistics here is basically adapted from Weir and Cockerham (1984), which is in case of F_{ST} a method of moment estimate with bias correction. To calculate the three F-statistics three variance components are needed, considering a single biallelic locus. The F_{ST} statistic bases on the ratio of this three variance components. $\bar{n}$ is the average sample size and N the sum of the n_i with $i = 1, \ldots, I$. The squared coefficient of variation of sample size is formulated by nc:

$$nc = \frac{(I\bar{n} - N^2)/(I\bar{n})}{I - 1}$$

$\bar{p}$ is the average sample frequency of the dominant allele p_i:

$$\bar{p} = \sum_{i=1}^{I} \frac{n_i p_i}{I\bar{n}}$$

The sample variance of the frequency of allele p_i over all populations is described by s:

$$s = \sum_{i=1}^{I} \frac{n_i (p_i - \bar{p})^2}{(I-1)\bar{n}}$$

and the average heterozygote frequency of allele p is represented by $\bar{h}$:

$$\bar{h} = \sum_{i=1}^{I} \frac{n_i h_i}{I\bar{n}}$$

Now that the variance terms can be calculated, the first is the variance between populations:

$$a = \frac{\bar{n}}{nc}\left(s - \left(\frac{1}{n_i - 1}\right)\left(\bar{p}(1-\bar{p}) - \left(\frac{I-1}{I}\right)s - \left(\frac{1}{4}\bar{h}\right)\right)\right)$$

where $1/4\bar{h}$ represents the effect of dominance deviation. The second is b, the variance between individuals within populations:

$$b = \left(\frac{\bar{n}}{\bar{n} - 1}\right)\left(\bar{p}(1-\bar{p}) - \left(\frac{I-1}{I}\right)s - \left(\frac{2\bar{n} - 1}{4\bar{n}}\right)\bar{h}\right)$$

The third, c, is the variance between gametes within individuals:

$$c = \frac{1}{2}\bar{h}$$

With the variance terms the F_{ST} statistics can be written shortly as shown in Equations 4 to 6. Not available estimates or also missings for F_{ST} can occur if a+b+c become null, which happens if all individuals of both populations have all the homozygous dominant

or homozygous recessive allele or are all heterozygous. To sum up, a single F_{ST} value is an estimate of differentiation of two populations, based on allele frequency and is used for the further analysis. The R-Code is shown in Section A.1.4.

$$\theta = F_{ST} = \frac{a}{a+b+c} \tag{4}$$

$$F = F_{IT} = \frac{1-c}{a+b+c} \tag{5}$$

$$f = F_{IS} = \frac{1-c}{b+c} \tag{6}$$

3 Comparison of two populations at a single locus

With the parameters of interest it is now possible to separate populations. The methods of the statistical testing, used in this work to separate two populations, are explained in this chapter.

3.1 Parameter estimiation

3.1.1 Allele frequency

The populations are compared by the estimated allele frequencies $\hat{h_i}$ and $\hat{p_i}$ of the sampled individuals. The populations size n_i is counted for each population. Individuals can either have an allelic state measured by the genetic marker of e.g. 0, 1 or 2 which indicates either the homozygous dominant, heterozygous or homozygous recessive genotype.

3.1.2 F_{ST}

As described, the F_{ST} statistic, represented by the parameter θ, is estimated by $\hat{\theta}$ for the comparison of two populations based on the allele frequencies and size of each population.

3.2 Hypotheses

The null hypothesis, required for statistical testing, is tested at a given type I error rate α. The null hypothesis is rejected if the θ value is greater than null.

$$H_0 : \theta = 0$$

$$H_A : \theta > 0$$

The H_A is accepted at the given α, because the H_0 can be rejected, because $\theta > 0$. In this case a difference between the populations is shown.

3.3 Error rate

The result that H_0 is rejected, if in truth the populations are equal, should occur maximal at α percent of the cases if the experiment is repeated.

3.4 Test statistic

The test statistic is simply

$$\hat{T} \;=\; \hat{\theta}_O$$

where $\hat{\theta}_O$ is the 'original' test statistic of the not resampled values, as shown in Equation 4. θ_O is used directly and the statistic is not standardised over the variance, as e.g. in a t-test.

3.5 Distribution of the test statistic

Bootstrapping is a statistical method of resampling for statistical inference and was first described by Efron (1979) and can be used to estimate the accuracy of statistical estimates, e.g. their the standard error, bias or mean. The bootstrap here is used to get an empirical F_{ST} distribution (Efron and Tibshirani, 1993). This distribution bases on the used data. Resampling usually means that the observed values are completely rearranged at random to the populations with repetition if the aim is statistical testing. After that, the location of θ_O in the resampled distribution is declared to be under the null hypothesis or not. The bootstrap here does not completely rearrange the individuals at random to the populations, but is stratified for each population. A nonparametric bootstrap procedure is used with a stratification by subpopulations. At a nonparametric bootstrap the theoretical distribution function is replaced by a nonparametric estimate and a simulated empirical distribution is the result (Efron and Tibshirani, 1993; Büning and Trenkler, 1998). For a given sample, repeated samples are drawn with replacement of the same size as the original. Some values of the original sample are not in the bootstrap sample, some can therefore be included once or more. The resampling should always be under the null hypothesis. To get the unknown F_{ST} distribution under H_0, populations with an equal frequency of the dominant allele are simulated and bootstrapped to get example data. In each of the B bootstrap replications with $b = 1, \ldots, B$, a bootstrap F_{ST} statistic is calculated by giving each population values. That are the allelic states 0, 1 or 2 of the individual haplotypes, of the original observed data, with a population index for each individual. The distribution is essential for statistical testing, because out of it a p-value can be calculated which is used to test the null hypotheses. The distribution bases on the results of

$$\hat{\theta}_b - \;\hat{\theta}_O$$

and is estimated of each of the B bootstraps where $\hat{\theta}_O$ is the original test statistic and $\hat{\theta}_b$ the bootstrap test statistic. θ was defined by Wright (1951) to range between null and one. Through this simple estimation of the distribution under H_0 a negative θ, which is not defined, can occur. The needed quantile for the p-value is the original test statistic $\hat{\theta}_O$ and is at the upper side of the distribution. It is a cut point at which the H_0 is rejected and H_A is valid.
Another possibility instead of the bootstrap distribution would be the assumption of a distribution to the test statistic. A χ^2 or Dirichlet (Weir and Hill, 2002) distribution are an alternative, because θ has quadratic terms. These distributions are adequate assumptions but probably at high sample sizes only. The better the assumed distribution fits to test test statistic, the better are the results. That is because it is used to estimate p-values, confidence intervals, bias and variance of an estimator for the statistic of interest (Efron, 1979; Westfall and Young, 1993). As mentioned above $\hat{\theta}_O$ is used to decide for H_0 or H_A. There are of course other ways of getting this cut point as there are options of using a different distribution as the H_0 bootstrap distribution but the bootstrap is one of the easiest to implement and should work (Davison and Hinkley, 1997; Efron and Tibshirani, 1993). As an example, bootstrapping over loci was performed by Weicker et al. (2001) with 5000 replications for estimates of Cockerham's θ with upper and lower limits but not stratified over populations. The result of Weicker et al. (2001) is that the estimated $\hat{\theta}$ are quite close to the actual θ, depending on the sample size. The larger the populations are the smaller is the difference between the estimated and the actual θ.

3.6 p-value

Comparing populations by F_{ST} requires a cut point at which populations can be declared to be different but there do not exist any clear F_{ST} constraints (Charlesworth, 1998). A solution for that is the p-value. The p-value is the probability of sampling observations, to be more extreme than the actual sample, given the H_0 to be true. For the p-value here at first an estimate of the original statistic $\hat{\theta}_O$ of a comparison between two populations and the H_0 distribution out of the bootstrap are needed.
The p-value is

$$\hat{p} \;=\; \frac{\#\ (\hat{\theta}_b > \hat{\theta}_O)}{B}$$

where # is the number of greater bootstrap test statistics and $\hat{\theta}_b$ the estimated resampled bootstrap test statistic. The p-value is then the proportion of greater $\hat{\theta}_b$ to $\hat{\theta_O}$ under the mentioned model- and distributional assumptions. In this work it is intended to compare two populations by a p-value based on $\hat{\theta}_O$ and calculated by a bootstrap to reject H_0 with a certain α. If the H_0 is rejected the H_A, the populations are different, is assumed.

3.7 Discussion

The used method to compare two populations has a drawback. The statistic is not standardised over the variance, as for example the t-test statistic. The variances of the test statistic are not estimated, because it would be quite extensive and probably not much more satisfying as it is now. A sample variance is included in the calculation of the F_{ST} statistic but no variance of F_{ST} is estimated. This is not optimal but very simple. Weir and Cockerham (1984) suggest the use of a jackknife procedure to estimate sample variances. Jackknifing is a resampling method as the bootstrap but with the difference that for each repetition a data point or observation out of the original data is omitted.
Of course the null hypothesis can not be proved. If no difference between the populations can be shown, it could be that the sample size is too small, the allele frequencies are too low or the distribution is too discrete. The discreteness makes it difficult to get a proper distribution under H_0, so that a cut point can not be estimated properly.

4 Global comparison of multiple populations at a single locus

The next step, after the comparison of two populations is explained, is the comparison of multiple populations. The important fact of the global comparison is the control of a global error rate for all comparisons simultaneously. All populations are compared together. This can then have the result, that there no difference between the populations can be shown. The alternative is that there is a difference between them. This difference can be due to one or more populations but it is not known which populations can be separated.

4.1 Parameter estimation

4.1.1 Allele frequency

For the I populations I values for each n_i, $\hat{h_i}$ and $\hat{p_i}$ have to be estimated, each as described for two populations.

4.1.2 F_{ST}

For the I populations a single global $\hat{\theta}$ is calculated out of n_i, $\hat{h_i}$ and $\hat{p_i}$ which is applied for all populations.

4.2 Hypotheses

The global null- and alternative hypothesis over all of the multiple populations looks similar to the hypotheses for two populations but with a different parameter used.

$$H_0 : \ \theta_G = 0$$

$$H_A : \ \theta_G > 0$$

H_0 can not be rejected, if the global θ_G is null. H_A is assumed if at least due to one of the populations θ_G becomes greater than null.

4.3 Error rate

The error rate at the global hypothesis is α as in the two population comparison, because only a single hypothesis is tested. In maximal α percent of the repetitions H_0 is rejected although it is true, if the experiment is repeated.

4.4 Test statistic

The test statistic is again simply

$$\hat{T}_{Glob} \ = \ \hat{\theta}_O$$

because only a single statistic is estimated for all populations. The calculation of θ_O is shown in Equation 4.

4.5 Distribution of the test statistic

Similar as for two populations, the distribution under the null comes from

$$\hat{\theta}_b \ - \ \hat{\theta}_O$$

and is calculated for each of the B bootstraps but with the global theta is used.

4.6 p-value

The p-value for the global hypothesis is calculated similar as for two populations but of course the global θ is used.

$$\hat{p}_G \ = \ \frac{\#\ (\hat{\theta}_b > \hat{\theta}_O)}{B}$$

4.7 Discussion

The use of a global hypothesis for multiple populations is the simplest step forward, if more than two populations are considered. Because only a single θ is calculated, the result is as easy to interpret as the results for two populations. If the H_0 can not be rejected this is a sufficient result for more than two populations. The result of a global test is a bit similar to the result of an ANOVA. It has the advantage of comparing multiple populations without concerning about adjustments for multiple hypotheses. The disadvantage, when H_0 is rejected and H_A accepted, is that the result can not be sufficient for the experimenter, because mostly it is of interest due to which population(s) H_0 is rejected. It stays unclear how many and which of the populations can be separated by θ. If this is of interest, an all pair comparison between the populations can be done.

5 All pair comparison of multiple populations at a single locus

If no difference between the populations can be shown, the all pair comparison does not deliver further information. It delivers further information if there is a difference between the populations, because it shows the differing populations.

5.1 Parameter estimation

5.1.1 Allele frequency

For the I populations I values for each n_i, $\hat{h_i}$ and $\hat{p_i}$ with $i = 1, \ldots, I$ have to be estimated, as described for two populations.

5.1.2 F_{ST}

For I populations, $\hat{\theta}_{Oj}$ values are calculated with $j = 1, \ldots, J$ and $J = I(I-1)/2$.

5.2 Hypotheses

With more than two populations an all pair comparison should be done. The hypotheses cover all of the j local hypotheses with $J = I(I - 1/2)$. The decision to reject the H_0 bases on $\hat{\theta}_j$. The H_0 is rejected if the corresponding test statistic is greater then null. H_0 can not be rejected if the corresponding comparison has a $\hat{\theta}$ of null. If at least one comparison has a θ greater than null, then the H_0 has to be rejected and the alternative is assumed.

$$H_0 : \bigcap_{j=1}^{J} \theta_j \; = \; 0$$

$$H_A : \max \theta_j > 0$$

In H_A the maximal of θ_j is used. If it is not greater than null, all other H_A can not be accepted.

5.3 Error rate

At more than two comparisons, or more precisely multiple hypotheses, different error rates can be defined. The local error rate covers a single hypothesis. There are as much local error rates, as there are hypotheses and each true H_0 hypothesis is rejected at most with the probability α. The global error rate has at most the probability of α of rejecting a true H_0 under the assumption, that all H_0 are true. The family wise error rate (FWER) belongs to the multiple error rate and is the probability α for rejecting at least one H_0, given the H_0 is true (Horn and Vollandt, 1995). The FWER should be controlled but there are two types. The weakly FWER is controlled if all H_0 are true. Even if a single H_0 is rejected, the global error rate is no more controlled for the other hypotheses. In the strong sense the FWER is controlled independently of true or false null hypotheses. This is desired if multiple hypotheses are of interest (Horn and Vollandt, 1995). The multiplicity adjusted pairwise F_{ST} test statistics should keep the FWER, which is shown in Section 6.5.

5.4 Test statistics

The test statistics are

$$\hat{T}_{AP} = \hat{\theta}_{bj} - \hat{\theta}_{Oj}$$

with $b = 1, \ldots, B$ and $j = 1, \ldots, J$.

5.5 Joint distribution of the test statistics

The difference between the bootstrap for two populations and the bootstrap for multiple hypotheses is, that the joint distribution of the test statistics is of interest instead of a single distribution. There are more than one distribution of test statistics. The test statistics are correlated because of the all pair comparison, e.g. 1 vs. 2 and 2 vs. 3. Therefore the maximal test statistic method is used. There are J comparisons and therefore J test statistics. As described by (Westfall and Young, 1993), but for a different statistic, the maximal bootstrap test statistic of θ_{bj} is taken and every statistic which is greater than the original, is counted. For each of the J original test statistics at each bootstrap replication it is counted, if any of the J bootstrap test statistics θ_b is greater than one of the J θ_O. This leads to a local conclusion under a controlled FWER. If a single θ is greater than null the H_0 is rejected, which than counts for all smaller F_{ST} statistics, too.

If there are more F_{ST} statistics, then a multivariate distribution can be assumed which means J dimensions instead of a univariate distribution if two populations with a single hypothesis are tested. The more dimensions there are, the harder it becomes to get a cut point, because the more populations there are the more has to be cut of the area where the H_0 can be rejected. In Section 6.4 this is illustrated for three populations but for two of three comparisons only.

5.6 Adjusted p-value

If more than one hypothesis is considered, then through the Bonferroni inequality the probability of rejecting H_0 although it is true increases. The multiple α raises with the number of hypotheses. This is solved with an adjusted p-value by using the maximum of the multiple test statistics. The p-value is the proportion of the maximal bootstrap θ_{bj} statistics greater than the original statistic θ_{Oj}, divided by the number of bootstraps B. The p-value is calculated under the assumption of a multivariate distribution, because of the multiple comparisons and is now already adjusted for this multiplicity

$$\hat{p}_j \;=\; \frac{\#(max\ \hat{\theta}_{bj} \;>\; \hat{\theta}_{Oj})}{B}$$

with $b = 1, \ldots, B$ and $j = 1, \ldots, J$. As a mathematical definition the multiple p-value is:

$$\widehat{p_j} \;=\; inf\ \{\alpha \mid H_j\ is\ rejected\ at\ FWE\ = \alpha\}$$

Where H_j are the null hypothesis, $\widehat{p_j}$ is the adjusted p-value or also the smallest significance level, where H_j can be rejected, using a simultaneous test procedure.

5.7 Discussion

Westfall and Young (1993) also took the maximum of the test statistics with the assumption of a multivariate distribution but resampled residuals in a linear model. They standardised over the variance, but in this work the variance of $\hat{\theta}_O$ under H_0 is not estimated. If the variance of F_{ST} would be estimated, the used method would be more similar to Westfall and Young (1993) but still with a different distribution assumed. Where here the distribution is unknown and empirical, Westfall and Young (1993) assumed normally distributed observations in a linear model with resampling to make conclusions about the distribution of the residuals, so the conclusions are not directly negotiable. The advantage of using the maximum of the bootstrap test statistics is that the FWER is controlled. It was not tested, if it is controlled strongly or weakly but the results are quite conservative. The control of the FWER strongly is desirable for multiple comparisons. There are other options as e.g the Bonferroni α adjustment, which divides α by the number of hypotheses.

The disadvantage of the Bonferroni adjustment is, that it becomes conservative with an increasing number of hypotheses (Horn and Vollandt, 1995). The other disadvantage is, that uncorrelated test statistics are assumed, which is not the case at the used all pair comparison of populations.

6 Simulations

6.1 Generation of Populations

As described, genetic markers are distinct and they are used at sampled individuals of populations to separate the populations. That marker data contains information to explain the variability within and between the populations. The 'ideal' marker should be as much informative as possible, this excludes missing values and includes codominance. The representative individuals of the populations are the basis for marker data. For simulations the populations are generated under this assumptions of an ideal marker, e.g. to ensure that no missing values occur and that the true allele frequencies are known. Populations can be described by the frequency of the dominant allele (p) per locus and the population size. The populations for all simulations are based on the HWE so each genotype frequency of the dominant allele p is calculated under the HWE. The allele frequency p is a single value of a single locus of each population. Each population has discrete allele frequencies. The used population sizes are equal, but it is also possible to create different population sizes. One way to summarise the individuals of a population is shown in Figure 3. There are three populations with ten individuals each and no missing value. Each field contains a value of 0, 1 or 2. This stands for each genotype or marker condition. Null would be the homozygous recessive genotype aa, one the heterozygous Aa and two the homozygous dominant genotype. This information can come for example from a polyacrylamid gel electrophoresis of a PCR with null, one and two marker bands. The values 0, 1 and 2 are drawn from the multinomial distribution. This distribution is chosen, because it is the multivariate generalisation of a binomial distribution and, instead of the two values null and one, three endpoints can occur. The R code, to generate the populations as described, is shown in Section A.1.1.

	Pop1	Pop2	Pop3
1	0	2	1
2	0	2	2
3	0	0	2
4	0	1	2
5	1	2	2
6	0	1	2
7	0	0	1
8	0	1	2
9	0	1	2
10	0	1	2

Figure 3: Example output of the function to generate populations. Three populations are shown with ten individuals each. The given frequency of p for the populations one to three are 0.1, 0.5 and 0.9.

6.2 Bootstrap distribution of F_{ST} under H_0 for two populations

6.2.1 Settings

For the bootstrap distribution two populations with an equal sample size of 100 are compared. 1000 bootstraps are done. The simulated allele frequencies of p_i are 0.001, 0.0025, 0.005, 0.01, 0.025, 0.05, 0.1, 0.2, 0.3, 0.4 and 0.5.

6.2.2 Results

The density of the distributions of the F_{ST} statistic given the allele frequency of p_i with 0.001, 0.0025, 0.005, 0.01, 0.025 and 0.05 are shown in Figure 4. For each p_i new populations are generated, as described randomly out of a multinomial distribution. Bootstraps are done and the distribution of the bootstrap test statistics is shown. The percentage of not available estimates to test statistics, for the allele frequency of this particular simulated populations, is 35.5% for 0.001, 36.4% for 0.0025, 14.2% for 0.005, 1.3% for 0.01 and null for the rest. This not available estimates can also be called missings. Missings are set to an F_{ST} statistic of null, because this samples do not contain information to differentiate the populations. The histograms show the density of bootsrap F_{ST} statistics so the histogram has a total area of one. The higher the density, the more bootstrap test statistics have this certain value. This varies with the generated population and the allele frequencies. The distribution of the frequencies 0.1, 0.2, 0.3, 0.4 and 0.5 are shown in Figure 8 in the appendix. There is also an example with an allele frequency p of 0.5 but with 1000 individuals per population in Figure 9 which has a shape, that is not so influenced by the population size or by the minimal allele frequency. The distribution will look similarly, if it repeated with a new population but the same settings.

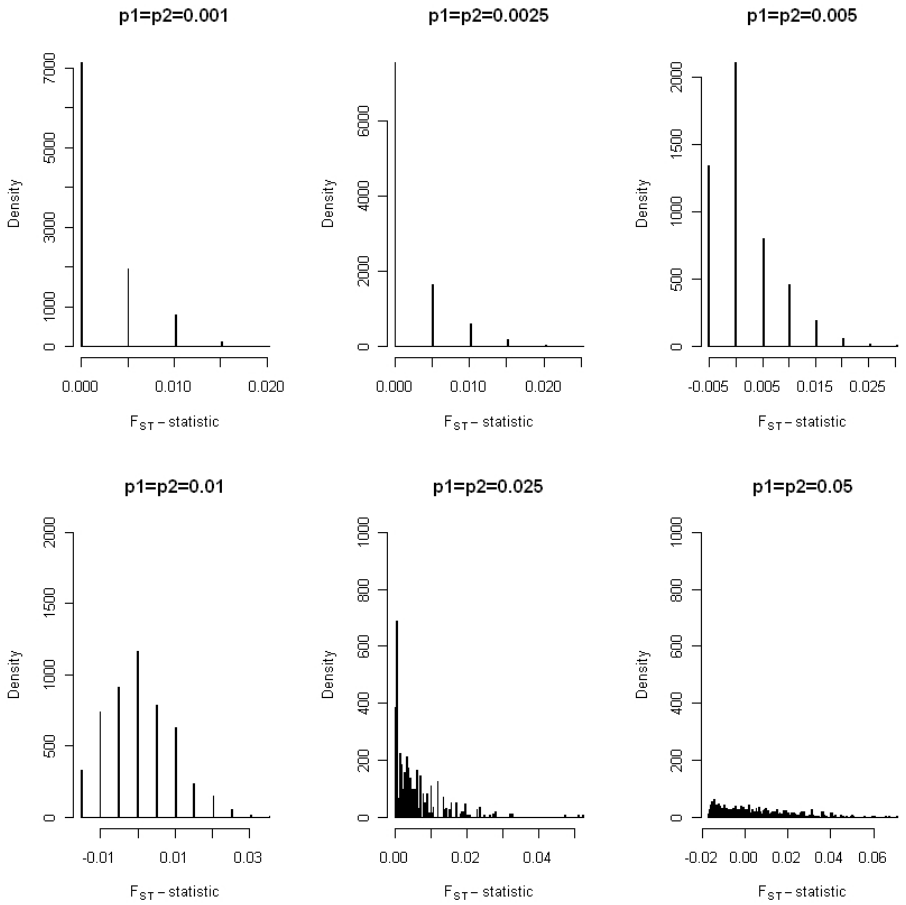

Figure 4: Simulated F_{ST} density distributions from 1000 bootstraps for the allele frequency p_i with 0.001, 0.0025, 0.005, 0.01, 0.025 and 0.05 for both of the two randomly generated populations with 100 individuals. Missings are set to null.

6.2.3 Discussion

Missings occur if all individuals of both populations have all the homozygous dominant or homozygous recessive allele or are all heterozygous. It is assumed that in this case there is no difference between the populations, because there is no information and the statistic is set to null, which can then be shown in the histogram. There is a higher density at $\theta = 0$, due to the missings. Another possibility would be, to consider only calculated results and ignore not available estimates. The reason for missings is known and the conclusion, to decide in these cases to not reject H_0, seems more plausible than the alternative. The shape of the density distribution as the proportion of missings varies with the generated marker. If a new population is drawn of which the bootstraps come from, the shown results will be different especially at small population sizes and extreme allele frequencies. The percentage of missings for p = 0.001 is one time 100%, where at other generated data it is only 14.3% and also the profile is different. With more individuals the distribution becomes more invariant, as shown in the example of 1000 individuals in

Figure 9 in the appendix. The closer the allele frequency gets to the boundary of null or one, the greater is the possibility of missings. The less the difference between the populations is, the higher is the discreteness of the distribution which can also be a single value. A small sample size can lead to a close estimate of the allele frequency and therefore a discrete F_{ST} distribution, too. A discrete distribution makes it difficult to estimate the p-value accurately, because it is either too conservative or too liberal. A population size of 100 individuals is chosen for this simulations to avoid too discrete distributions at all tested allele frequencies and because it is not too far away from practice. The bootstrap distribution can get negative as shown e.g. at allele frequency p = 0.05 and is therefore a bit less discrete. This is partly because of θ_b - θ_O but also because of the bootstrap variation of θ_b.

6.3 Global bootstrap distribution of F_{ST} under H_0 for multiple populations

6.3.1 Settings

For the global bootstrap distribution of six populations with an equal sample size of 100, 1000 bootstraps are done. The simulated allele frequencies of p_i are 0.001, 0.0025, 0.005, 0.01, 0.025, 0.05, 0.1, 0.2, 0.3, 0.4 and 0.5.

6.3.2 Results

The density of the distributions of the global F_{ST} statistic, given the allele frequency of p_i with 0.001, 0.0025, 0.005, 0.01 and 0.025, are shown in Figure 5. For each p_i new populations are generated randomly, bootstraps are done and the distribution of the bootstrap test statistics is shown as a histogram. The percentage of missings to the number of test statistics for the allele frequencies is 34.6% for 0.001, 2.1% for 0.0025, 4.4% for 0.005 and null for the rest. The missings are set to an F_{ST} statistic of null. The maximum density decreases the closer the allele frequencies are to 0.5. The distributions of the allele frequencies 0.05, 0.1, 0.2, 0.3, 0.4 and 0.5 are shown in Figure 10 in the appendix. There is also a more invariant example with an allele frequency p for all populations of 0.5, because 1000 individuals per population are used. This is shown in Figure 11 in the appendix.

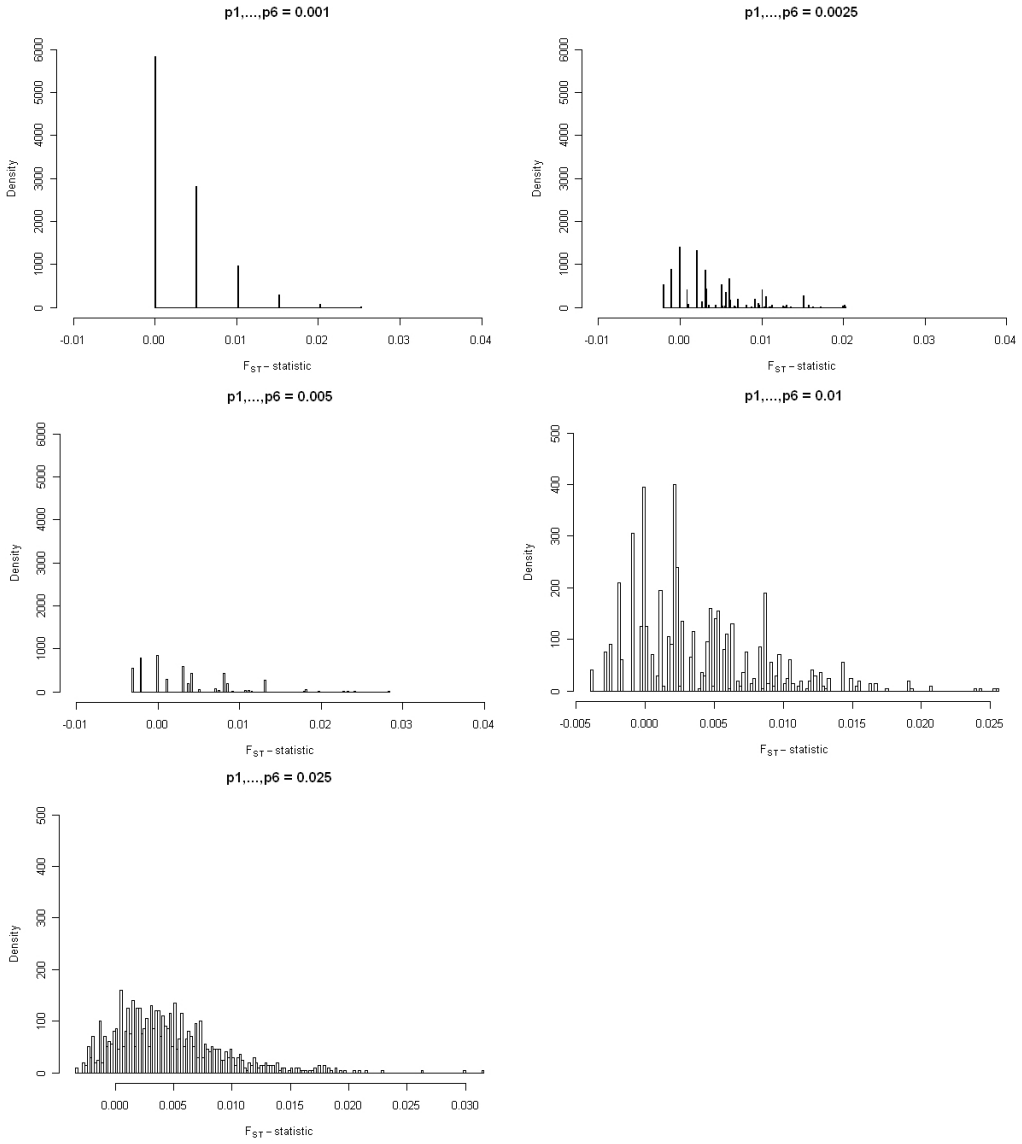

Figure 5: Simulated global F_{ST} density distributions of six populations from 1000 bootstraps for the allele frequency p_i with 0.001, 0.0025, 0.005, 0.01, 0.025 for all of the randomly generated populations with 100 individuals. Missings are set to null.

6.3.3 Discussion

In a reduced form the results of the global bootstrap distributions correspond to the results for two populations. However, they are not comparable to each other, because the basis e.g. the degree of freedom and so the distributions are different. The F_{ST} value is calculated from six populations. All of them must have identical information about the individuals, so that a missing can occur. The more populations there are, the less missings should occur, because of the higher sample size. The discreteness diminishes with decreasing extreme allele frequencies. The skewness of the global F_{ST} looks a bit similar to an F-distribution with df > 2, except the smallest frequency of 0.001. It is a quadratic statistic and the resulting skewness is affected by the F_{ST} adjustment but it

can not eliminated. There many of reasons that affect the skewness of the distribution. One of them is, that the variance of an allele frequency at the border of the frequency range can not be as high as an allele frequency in the middle of the frequency range. The higher the population size is, as e.g. shown with 1000 individuals instead of 100, the less is the difference in the shape of the distribution if new populations are created and a new bootstrap distribution is shown.

6.4 Bootstrap distribution of F_{ST} values of an all pair comparison for multiple populations under H_0

6.4.1 Settings

For the multivariate bootstrap distribution three populations with an equal sample size of 100 are compared and 1000 bootstraps are done. The simulated allele frequencies of p_i are 0.001, 0.0025, 0.005, 0.01, 0.025, 0.05, 0.1, 0.2, 0.3, 0.4 and 0.5.

6.4.2 Results

The results of the multivariate bootstrap density distribution, given the allele frequencies of 0.0025, 0.005, 0.01 0.025, 0.05 and 0.1, are shown in Figure 6 and the frequencies from 0.2 to 0.5 in Figure 12 in the appendix. For each p_i new populations are generated at random, bootstraps are done and the distribution of the bootstrap test statistics is shown. The plots show the correlation of two comparisons, where each includes an identical population, here 'Pop1'. Pop2-Pop1 is on the ordinates and Pop3-Pop1 on the abscissas. The boundaries should mark the area of negative F_{ST} values. For the first setting of $p_i = 0.001$ there are no estimates available so it is not shown. The number of missings for the other settings are up to 379 of the 1000 bootstraps. At $p_i = 0.0025$ there are 366 missings at the comparison 'Pop2 - Pop1' and 143 at 'Pop3 - Pop1'. At $p_i = 0.005$ there are 339 and 379 missings for the two comparisons, 2 and 49 for $p_i = 0.025$, respectively. At the other settings the number of missings is null. Not available estimates are set to null. The comparison of the results between the frequencies of 0.0025 to 0.1 shows the decrease of the discreteness, the closer the allele frequency is to 0.5. The difference in the discreteness between the frequencies 0.2 to 0.4 is not as high as at the smaller frequencies.

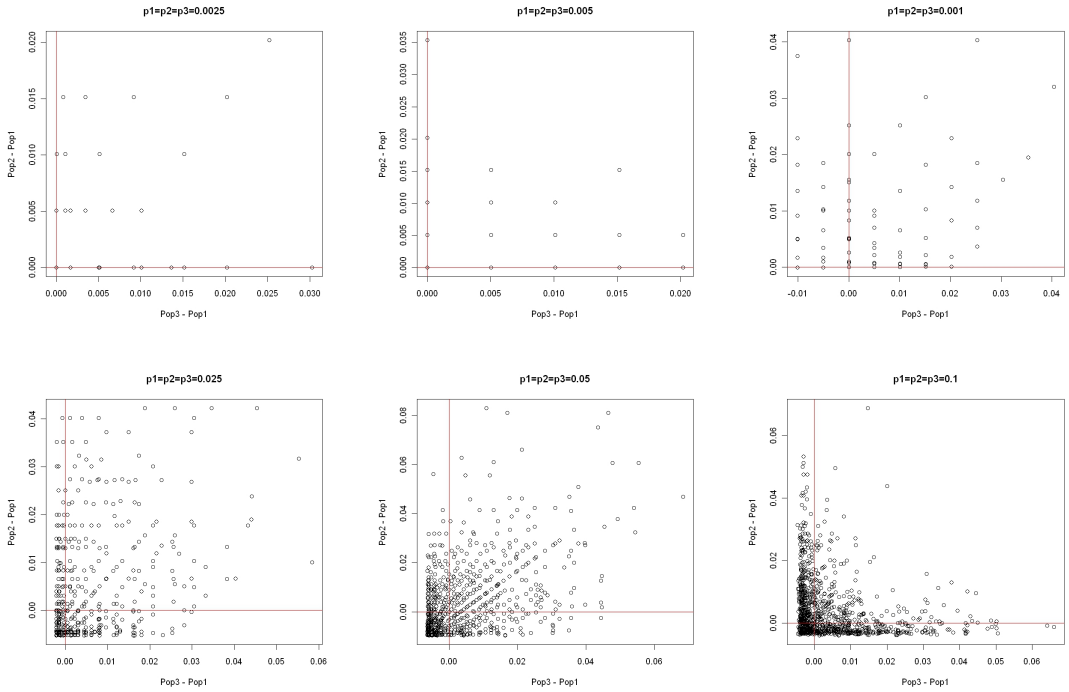

Figure 6: Simulated multidimensional F_{ST} density distributions. Shown are two of three of the all pair comparison. Three populations with 100 individuals, which are generated out of a multinomial distribution, are compared. They equal allele frequencies of p_i are set to 0.0025, 0.005, 0.01, 0.025, 0.05 or 0.1. 1000 bootstrap replications are done. Not available estimates are set to null.

6.4.3 Discussion

Illustrating a F_{ST} distribution of an all pair comparison can not be clearly arranged and becomes harder with the number of comparisons, because the number of dimensions increases. Therefore only the illustration of two comparisons is reasonable to be shown. As expected the distribution is discrete at a small allele frequency and becomes less discrete the closer the frequency is to p = 0.5. Estimates behind the boundaries, or to be more precise smaller than null, are in an undefined area. A value of null represents completely identical populations and a value smaller than null is not interpretable. The more populations are compared, the more dimensions occur and the more conservative the statistic becomes. The area which is smaller as the border θ_O shrinks with the number of dimensions, because the area under H_0 increases. The more dimensions there are, the more bootstrap repetitions are needed to estimate the distribution or the discreteness will increase.

6.5 α simulations

At the α simulations the populations are compared under H_0. This is done to see, if the test is liberal or conservative. The type I error rate α is then the number of com-

parisons with differentiated populations. The measured α is the indicator for a liberal or conservative test. If the actual α is higher as given, the test is liberal and the test is conservative if the actual α does not reach the given α. The simulations are done with an equal allele frequencies p_i for all populations, so that there is no difference between the populations. The test is allowed to reject H_0 maximal in $\alpha\%$ of the repetitions and lead to a false positive result, because H_0 is true but rejected. At multiple populations, more than one comparison is made. That increases the probability to see a difference between two populations, although in truth there is no difference. The more comparisons with a local α are made, the higher is the probability to get a false positive result and to exceed the given global respectively multiple α. To avoid this, a method is used, where the maximum test statistic of all comparisons is compared with the original test statistic. The FWER should therefore be controlled.

6.5.1 Settings

Two and six populations are compared with an allele frequency p_i of 0.5 at a single locus each. The simulated population sizes are 75, 100, 150, 200, 250, 500, 1000, 1500, 2000, 2500, 3000 and 3500 which gives a total of twelve settings. If six populations are compared, this is done at a global hypothesis and with an all pair comparison, too. The size of the test is computed for each setting with 10000 replications and 1000 bootstraps at each replication. The alpha is set to five percent for the local hypotheses as for the global and the family wise error rate.

6.5.2 Results

The results are shown as test size for each population size and population setting in Table 12 in the appendix. The α simulations show, that the type I error is lower as it is set. In most of the cases it is nearly null, only for two populations it is up to 2e-04 e.g. at 1500 individuals or 1e-04 at 2500 individuals at the global comparison. Even an increased population size has no effect to α, independent of the way the populations are compared.

6.5.3 Discussion

The F_{ST} statistic delivers conservative results, because in only under one percent of the cases the null hypothesis is rejected although it is true. The desired result would be, if $\alpha\%$ of the null hypothesis are rejected although they are true, because this type I error is set to be allowed. With the actual usage of α it can happen that H_0 is accepted too often even though there is a difference between the populations and H_A is true. One reason should be, that the empirical distribution of the test statistic under H_0 does not represent the true distribution, because the difference to θ_O is used. At multiple populations this simulation shows that the FWER is controlled, because the size of the test is not greater than the

set α. Westfall and Young (1993) developed the method of the maximum test statistic to control the FWER under different conditions. Here a nonnormal distribution is estimated instead of the valid assumption of an analytic distribution, as the studentized range would be and so the usage of α is of interest. The FWER in the strong or weakly sense can not be claimed, because it was not simulated but the conservative results indicate that they will be probably kept.

6.6 Two populations at a single locus

6.6.1 Settings

The parameter settings for the comparison of two population are a set of all combinations of the following variables. The allele frequencies of the dominant allele in both populations go from 0.05 to 0.95 in steps of 0.05. The population size of 5, 10, 25 and 50 is balanced. In total 1444 different settings are simulated. 10000 replications are done, so 10000 times new populations are created and each time 1000 bootstraps are done.

6.6.2 Results

In Figure 7 the results of the comparison between two populations are shown. For the four sample sizes the graphic has on the ordinate the allele frequency of p for population A and on the abscissa the frequency for population B. The graphic shows the power at each sample size where a different colour stands for a different range of power. The power is the probability of rejecting H_0 when H_A is true. Due to the settings the critical borders between the simulated 361 settings for each population size are done by interpolation and are in a range of 0.03 to 0.07. The results of the different allele frequency and sample size can be compared e.g. when a certain power is reached. The higher the sample size is, the less the difference of the allele frequency must be to reach the power of interest, which should be as high as possible under the alternative. Obviously the higher the difference of the allele frequency is, the better is the power to separate populations and also less individuals are needed to detect that difference. At a low sample size it is possible, that the certain power regions can not be reached. The power becomes better, the farther the allele frequency of the populations is of 0.5.

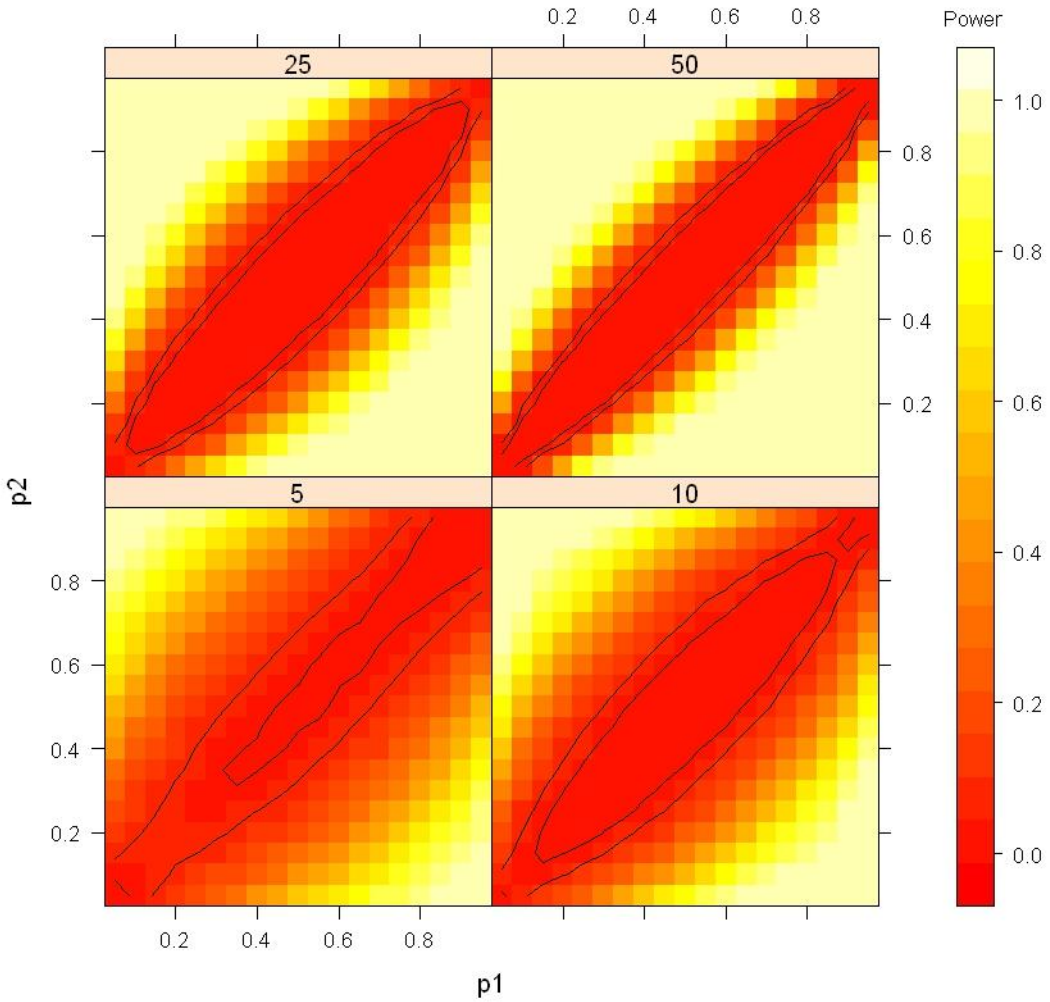

Figure 7: Allele frequency (p) of two populations and the corresponding power to detect a difference for four different population sizes.

6.6.3 Discussion

The ideal result would be a straight line with the preset error α of 0.05 if the allele frequencies are the same and a power of one if they differ. The higher the population size gets, the closer the result becomes the ideal result but it is not reached. Higher sample sizes could be simulated to get closer to the ideal result but this would not lead to further important conclusions. The used sample sizes are very common at experiments, where populations should be analysed by genetic marker data. A shrinkage of the difference of the allele frequency to get a power of 80% can be shown with an increasing sample size, which fullfills the prospects. The power curves are the steepest at the border allele frequency. This is comprehensible, because it is easier to show a difference the more opposed the populations are.

6.7 Global comparison of multiple populations at a single locus

6.7.1 Settings

The 14 settings for the four population sizes 5, 10, 25 and 50, that are tested, are shown in Table 3. At the first five settings, all of the six populations have an equal allele frequency p_i of 0.15, 0.25, 0.5, 0.75 and 0.85. The following four settings compare five populations with a frequency of 0.5 against one population with values for p of 0.15, 0.25, 0.75 and 0.85. These values of p for the aberrant population stay the same for the last five settings but the other five populations have an equal allele frequency of 0.1 instead of 0.5. Again the population sizes are balanced, 10000 replications and 1000 bootstraps are done, too. At the global comparison a single statistic is calculated for all populations at each setting.

Table 3: Settings of the dominant allele p for the global and all pair comparison of six populations, each for the population size of 5, 10, 25 and 50.

p setting	Pop1	Pop2	Pop3	Pop4	Pop5	Pop6
1	0.15	0.15	0.15	0.15	0.15	0.15
2	0.25	0.25	0.25	0.25	0.25	0.25
3	0.5	0.5	0.5	0.5	0.5	0.5
4	0.75	0.75	0.75	0.75	0.75	0.75
5	0.85	0.85	0.85	0.85	0.85	0.85
6	0.5	0.5	0.5	0.5	0.5	0.15
7	0.5	0.5	0.5	0.5	0.5	0.25
8	0.5	0.5	0.5	0.5	0.5	0.75
9	0.5	0.5	0.5	0.5	0.5	0.85
10	0.1	0.1	0.1	0.1	0.1	0.15
11	0.1	0.1	0.1	0.1	0.1	0.25
12	0.1	0.1	0.1	0.1	0.1	0.5
13	0.1	0.1	0.1	0.1	0.1	0.75
14	0.1	0.1	0.1	0.1	0.1	0.85

6.7.2 Results

The results of the global comparison of multiple populations are shown in Table 4. Table 10 shows the mean F_{ST} values in the appendix. In Table 4 the power to detect a difference between the populations is shown. The power is the probability, that a false null hypothesis will be rejected and varies between 0 and 1, so the range is completely used with the used settings. The power increases, as expected, with the sample size and with the difference between the populations. It has no influence on the power, in which direction the allele frequency of the aberrant population differs e.g. visibly at setting six

and nine. Non available power and F_{ST} estimates occur only at a population size of five. At this population size the settings one, ten and eleven showed one, six and four not available estimates. The results, if not available F_{ST} estimates are set to null and not available power estimates are set to one, are not influenced, because of the small amount of not available estimates. The range of the F_{ST} values, shown in Table 10, is between null and 0.5. As expected it is null if there is no difference between the populations and increases with the difference of the allele frequency. The influence of an increased sample size seems to be small, because the values are changed only slightly.

Table 4: Each of the 14 rows shows one setting of allele frequencies for each of the six populations and the corresponding power at a global comparison. Four different population sizes are tested.

Allele frequencies	n = 5	n = 10	n = 25	n = 50
0.15, 0.15, 0.15, 0.15, 0.15, 0.15	0.01	0.00	0.00	0.00
0.25, 0.25, 0.25, 0.25, 0.25, 0.25	0.01	0.00	0.00	0.00
0.5, 0.5, 0.5, 0.5, 0.5, 0.5	0.01	0.00	0.00	0.00
0.75, 0.75, 0.75, 0.75, 0.75, 0.75	0.01	0.00	0.00	0.00
0.85, 0.85, 0.85, 0.85, 0.85, 0.85	0.01	0.00	0.00	0.00
0.5, 0.5, 0.5, 0.5, 0.5, 0.15	0.06	0.11	0.57	0.98
0.5, 0.5, 0.5, 0.5, 0.5, 0.25	0.02	0.02	0.10	0.44
0.5, 0.5, 0.5, 0.5, 0.5, 0.75	0.02	0.02	0.10	0.44
0.5, 0.5, 0.5, 0.5, 0.5, 0.85	0.06	0.12	0.57	0.98
0.1, 0.1, 0.1, 0.1, 0.1, 0.15	0.01	0.00	0.00	0.00
0.1, 0.1, 0.1, 0.1, 0.1, 0.25	0.02	0.02	0.03	0.14
0.1, 0.1, 0.1, 0.1, 0.1, 0.5	0.21	0.38	0.90	1.00
0.1, 0.1, 0.1, 0.1, 0.1, 0.75	0.70	0.95	1.00	1.00
0.1, 0.1, 0.1, 0.1, 0.1, 0.85	0.89	1.00	1.00	1.00

6.7.3 Discussion

The results of the global comparison of multiple populations correspond to the expectations. The greater the difference of the allele frequency is, the higher is the power to show that the populations are differentiated and the higher is the simulated F_{ST} value. A smaller drawback is the not optimal usage of α, which is the same result as in the α simulations but for smaller sample sizes, and that no higher F_{ST} values than 0.5 occur. This can be explained by the hypothesis, that multiple populations are tested together and the effect of a single aberrant population is not as high if two different populations are compared. The number of not available estimates is smaller as in the comparison of two populations which bases on the same effect as before. A single F_{ST} value can

still be calculated if a comparison between two population does not deliver a result, because another comparison with a result leads to a result of the global hypothesis. A not available estimate at a global comparison occurs only, if all populations have individuals with identical genetic marker information and this probability decreases quickly at higher population sizes and less similar allele frequencies.

6.8 All pair comparison of multiple populations at a single locus

6.8.1 Settings

The 14 settings for the four simulated population sizes 5, 10, 25 and 50, are identical to the settings of the global comparison of six populations. They are shown in Table 3 and described in Section 6.7.1. Again the population sizes are balanced and 10000 replications with 1000 bootstraps are done, too. All pair comparisons are done, each population is compared with each other (Westfall and Young, 1993). For the first five settings the arithmetic average of the results of all comparisons is calculated and shown. At the next settings, all the contrasts against the population with the differing allele frequency are put together and this arithmetic average is shown.

6.8.2 Results

The results for the comparison of multiple populations are shown in the Tables 5, 6 and 11. The results of the F_{ST} value simulations are in Table 11 in the appendix, of the power simulations in Table 5 and the results of the proportion of not available estimates to available estimates are shown in Table 6. For each sample size the 14 different settings and their results are averaged and presented. The F_{ST} values are in a range between nearly null and 0.72. If every of the six populations has an equal allele frequency at the tested locus (0.15, 0.25, 0.5, 0.75 or 0.85), the results are almost null. Some results are minimally negative with -0.01. If five of the six populations have an allele frequency of 0.5 for p and the other population differs, the result changes symmetrical with the difference between the one population to the others. The frequencies 0.15 and 0.85 have nearly the same F_{ST} result and the frequency 0.25 and 0.75 are similar to each other, too. The last five settings show, that the statistic increases with the difference of the allele frequency between the five groups to the other. The results between the population size show a certain behavior, too. If every frequency is the same, the value is nearly null. With other settings an increased sample size leads to a minimal higher F_{ST} value.
The power to show a difference between the populations, the probability that a false null hypothesis will be rejected, varies between 0.01 and 1. It is as expected low with values nearly null or 0.01, when the frequencies are equal. The power increases with the sample size, if the frequencies are unequal. The power is nearly the same at the settings, where the distance to the populations with a frequency of 0.5 is the same, for example 0.15

respectively 0.85 vs. five times 0.5. The power rises also with an increasing difference of the allele frequency which is shown at the last five settings. The power is higher, if the allele frequencies of the five equal set populations are closer to the border of the frequency range. The power is not as high, if the five equal frequencies are in the middle, of course with nearly the same distance to the aberrant population, so that it is comparable. If a p-value could not be calculated it is set to one and the power is only as high as before or smaller. The more not available estimates there are, the more decreases the power. The percentage of not available estimates ranges between null percent and about 62%. They mostly turn up at the smallest sample size and at the end of the allele frequency range, where the equal population frequency 0.15 and 0.85 at a population size of five are a good example. There are always more not available estimates, approximately ten times, at the power data as at the F_{ST} data.

Table 5: Each of the 14 rows shows one setting of allele frequencies for each of the six populations and the corresponding power. An all pair comparison between the six populations is done. 10000 replications and 1000 bootstraps are done for each setting. The values in brackets are the power if not available estimates are turned into one. Four different population sizes are tested.

Allele frequencies	n = 5	n = 10	n = 25	n = 50
0.15, 0.15, 0.15, 0.15, 0.15, 0.15	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
0.25, 0.25, 0.25, 0.25, 0.25, 0.25	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
0.5, 0.5, 0.5, 0.5, 0.5, 0.5	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
0.75, 0.75, 0.75, 0.75, 0.75, 0.75	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
0.85, 0.85, 0.85, 0.85, 0.85, 0.85	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
0.5, 0.5, 0.5, 0.5, 0.5, 0.15	0.05 (0.05)	0.13 (0.13)	0.43 (0.43)	0.82 (0.82)
0.5, 0.5, 0.5, 0.5, 0.5, 0.25	0.02 (0.02)	0.03 (0.03)	0.08 (0.08)	0.22 (0.22)
0.5, 0.5, 0.5, 0.5, 0.5, 0.75	0.02 (0.02)	0.03 (0.03)	0.08 (0.08)	0.22 (0.22)
0.5, 0.5, 0.5, 0.5, 0.5, 0.85	0.05 (0.05)	0.13 (0.13)	0.44 (0.44)	0.82 (0.82)
0.1, 0.1, 0.1, 0.1, 0.1, 0.15	0.00 (0.00)	0.01 (0.00)	0.00 (0.00)	0.00 (0.00)
0.1, 0.1, 0.1, 0.1, 0.1, 0.25	0.02 (0.01)	0.04 (0.03)	0.05 (0.05)	0.09 (0.09)
0.1, 0.1, 0.1, 0.1, 0.1, 0.5	0.20 (0.09)	0.40 (0.35)	0.79 (0.79)	0.99 (0.99)
0.1, 0.1, 0.1, 0.1, 0.1, 0.75	0.64 (0.28)	0.92 (0.82)	1.00 (1.00)	1.00 (1.00)
0.1, 0.1, 0.1, 0.1, 0.1, 0.85	0.84 (0.36)	0.99 (0.88)	1.00 (1.00)	1.00 (1.00)

Table 6: Each of the 14 rows shows one setting of allele frequencies for each of the six populations and the corresponding not available estimates in % for the four different population sizes. The value without brackets comes from the power data, the value with brackets from the F_{ST} results. An all pair comparison between the six populations is done. Four different population sizes are tested.

Allele frequencies	n = 5	n = 10	n = 25	n = 50
0.15, 0.15, 0.15, 0.15, 0.15, 0.15	34.02 (3.98)	1.99 (0.14)	0.00 (0.00)	0.00 (0.00)
0.25, 0.25, 0.25, 0.25, 0.25, 0.25	4.29 (0.32)	0.02 (0.00)	0.00 (0.00)	0.00 (0.00)
0.5, 0.5, 0.5, 0.5, 0.5, 0.5	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
0.75, 0.75, 0.75, 0.75, 0.75, 0.75	4.48 (0.35)	0.02 (0.00)	0.00 (0.00)	0.00 (0.00)
0.85, 0.85, 0.85, 0.85, 0.85, 0.85	33.88 (3.87)	2.21 (0.13)	0.00 (0.00)	0.00 (0.00)
0.5, 0.5, 0.5, 0.5, 0.5, 0.15	0.11 (0.02)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
0.5, 0.5, 0.5, 0.5, 0.5, 0.25	0.03 (0.01)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
0.5, 0.5, 0.5, 0.5, 0.5, 0.75	0.02 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
0.5, 0.5, 0.5, 0.5, 0.5, 0.85	0.11 (0.02)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
0.1, 0.1, 0.1, 0.1, 0.1, 0.15	62.29 (6.77)	13.58 (0.48)	0.01 (0.00)	0.00 (0.00)
0.1, 0.1, 0.1, 0.1, 0.1, 0.25	58.19 (1.95)	11.33 (0.02)	0.05 (0.00)	0.00 (0.00)
0.1, 0.1, 0.1, 0.1, 0.1, 0.5	56.93 (0.06)	11.35 (0.00)	0.06 (0.00)	0.00 (0.00)
0.1, 0.1, 0.1, 0.1, 0.1, 0.75	57.01 (0.00)	11.31 (0.00)	0.02 (0.00)	0.00 (0.00)
0.1, 0.1, 0.1, 0.1, 0.1, 0.85	57.05 (0.00)	11.19 (0.00)	0.05 (0.00)	0.00 (0.00)

6.8.3 Discussion

The results show that an increasing population size leads to a higher power to detect a difference between the populations and to less not available estimates. Obviously, the greater the difference of the allele frequencies between the populations are, the better the populations can be separated. The power to show a difference between populations is high, if the allele frequencies of the five equal populations is at the border of the frequency range, and not as high, if the five the allele frequencies of the five equal populations are in the middle of the frequency range. This can be explained by the smaller variances at the border of the frequency range. The basis for this is shown in Section 2.2.2. The smaller the variances are, the better the populations can be separated. The effect that not available estimates are set to one at the power simulation has of course the greatest influence, where the most of them occur. This does not count for results, where the power is already nearly null e.g. at the first five settings with equal allele frequencies. Results as -0.01 can mainly be explained by the usage of θ_b - θ_O as the calculation for the bootstrap distribution. Next to that the effect of small sample sizes which lead to not available estimates. The power could not be calculated more often than F_{ST}, because the power bases on the F_{ST} distribution but if the F_{ST} statistic gets discrete it can still be calculated

mostly. The tendency of the results follows the expectations but is conservative as shown at the α simulations. A recommendation of a minimal sample size for this number of compared populations are ten individuals per population or sampling spot. Populations with less individuals can have a lot of missings and may not be representative for the populations. Furthermore the maximal power of 84% occurs at a extreme setting and is decimated if the missings are taken into account. Ten individuals seems quite small and it has the drawback, that at this sample size a practical power requires are huge difference of the allele frequencies between the populations. At ten individuals a difference of 0.65 results in a power of 82%. Even if it sounds trivial, the more individuals there are per population the better they can be separated. The smaller the difference between the populations is, the more important becomes a high sample size to show that difference.

7 Example data sets

Due to the simulations the characteristics and basics of the used test statistic are known and can be used to analyse some example data sets. The corresponding multiplicity adjusted and not multiplicity adjusted p-values of the pairwise comparisons for every example are shown in the Tables 14 to 19 in the appendix. Three examples are shown. Each of the results are from one example locus of the data set. At first the fugus data, the motivating example, is presented. Secondly cattle breeds are of interest in the microbov example. Thirdly a house cat data set is shown. The p-values are graphically displayed as letters. Equal letters represent that the populations are not significantly different, at a multiple α level of .05. The populations are significantly different, if they do not have any equal letters. The letters are generated with the function *multcompLetters* of the *multcompView* R-package, version 0.1-2 (Piepho, 2004).

7.1 Fungus data - Motivating example

The first example data set is the motivating example. The fungus data set is from a trial done by A.-K. Lühmann. As mentioned before, 16 possible loci could be analysed. Three are deleted due to technical reasons with the genetic markers. Out of the 13 still remaining loci the number of individuals with a missing value range between 70 at locus five and 192 at locus eight out of the total 716 sampled individuals. The used markers are codominant but are evaluated as dominant markers by the experimenter. The method, which is used so far to estimate the allele frequencies out of the data, is therefore changed. The frequency of the recessive genotype could be estimated directly as before. The frequencies of the heterozygous genotypes and the frequencies of the homozygote dominant genotypes are estimated under the assumption of the HWE. The basis is the measured frequency of the recessive allele. A bootstrap with 1000 replications is performed to calculate multiplicity

adjusted p-values out of the original data for an all pair comparison, as described in Section 5. For ten of the 13 loci no p-values could be estimated, because the markers are probably not polymorph. If they could be estimated, the number of p-values smaller than .05 is seven, nine and 17 out of 66 possible comparisons for Locus7, Locus4 and Locus16. These three loci do have in common, that the estimated frequency of the heterozygous genotype is relatively high, where the other loci do have a high amount of homozygous genotypes. After this description of the data set, Locus4 is used as an example for the further analysis. At this locus populations can be separated due to significant comparisons and at this locus there are only 71 values less than possible in the data set. The frequency of the homozygote recessive genotype over all populations, could be estimated directly from the data set. The rounded value is 0.57. Out of this the frequency of the heterozygous individuals and the dominant genotype must be estimated by the HWE. They rounded values are 0.37 and 0.01, respectively. There are also locations, with generally only few observations as e.g. 'Lauenau' with five and 'Kuesten' with three observed individuals. As recommended in the discussion of the simulations, the populations with this sample sizes are not taken into account. The assumption, that this amount of individuals is representative for the population, is probably not given. The population sizes vary between 14 for 'UNI' and 107 for 'Kordes'. The results of the all pair comparison of the remaining populations are shown in Table 7. The locations 'UNI' and 'Ruthe' are experimental fields, all other locations are breeding fields. The experimental fields can have a higher fungus heterogeneity due to fungi of other plants and trials. The breedings field should be more homogenic. The exclusion of the two locations with too less individuals effects the results. If they are included, less populations can be separated. This has several reasons. The more locations are tested, the more comparisons there are. The number of dimensions increases but not the number of bootstraps. The more dimensions there are the more bootstraps must be done to cover the area of the multidimensional distribution. The less values there are in this area the harder it is to estimate a p-value, because the cut point is probably not correct due to the discrete distribution.

Table 7: All pair comparison of the 13 different fungus sampling spots. Equal letters show that a significant difference between the fungi, at an α level of .05, could not be found. The results are from Locus4.

Fungus sample	Letters of no significance
Dortmund	abc
GGH	ab
Kassel	ab
Kordes	ab
Noack	a
Ruthe	bc
Sangerhausen	c
StHannover	ab
Tantau	abc
UNI	a
UNISORTEN	ab
Bremen	abc

7.2 Microbov

A good example is the microbov data set by Laloe et al. (2007) from the *adegenet* R-package. In this data set, the cattle breeds should be separated. The 704 cattle individuals either come from France or not further specified from Africa and belong to the species *Bos taurus* or *Bos indicus*. The data set contains 30 loci with corresponding microsatellite markers and for each of the loci up to 22 alleles. The used microsatellite markers are recommended by the Food and Agriculture Organization (Laloe et al., 2007). Due to the loci and alleles, 373 points can be analysed. Here the second allele of the third loci 'L03.02' is used as a example. The locus has rounded frequencies for p^2, 2pq and q^2 of 0.626, 0.298 and 0.075. An extract of the data set is shown in Table 13 in the appendix. Out of the used codominant marker, the genotype frequencies are estimated and under assumption of the HWE the allele frequencies of the populations, too. Six individuals of the 704 analysed, have no information about the allelic state and are omitted. For every individual, the data set provides information concerning the 15 breed types that are possible. The individuals should not be related among each other to avoid pseudo replications. Because of that, the individuals are rejected, if a related cattle is already measured. A bootstrap with 1000 replications is performed to calculate multiplicity adjusted p-values out of the original data for an all pair comparison, as described in Section 5. The results are shown in Table 8. Theoretically there are as much differentiations possible as there are measured breeds. Seven groups are made out of these 15 breeds, again indicated by letters, with an error of .05. Other loci would have a different result. The different group sizes of the 15 breeds of 29 to 61, are taken into account. Only 85 of the 373 loci deliver p-values, what leads to 288 loci where the calculation of a p-value is a problem. The maximum of p-values smaller than .05 at one locus is 50, where the loci L03.02 has 15 out of 105 possible comparisons.

Table 8: All pair comparison of the 15 cattle breeds. Equal letters show that a significant difference between the breeds at an α level of .05 could not be found. The results are from locus 'L03.02'.

Cattle breed	Letters of no significance
Bazadais	a
BlondeAquitaine	abcd
Borgou	abcd
BretPieNoire	abcd
Charolais	abcd
Gascon	abcd
Lagunaire	b
Limousin	a
MaineAnjou	bcd
Montbeliard	acd
NDama	abcd
Salers	a
Somba	a
Zebu	bc
Aubrac	cd

7.3 Nancycats

Another example from the *adegenet* R-package is the data set nancycats of the unpublished paper by Devillard et al. (2011). In this data set, house cat colonies are being separated. Nancycats has the genotypes of 237 cats (*Felis catus*) from 17 colonies. The colony size differs between nine and 23 and is taken into account. It contains nine loci with microsatellite markers and up to 18 alleles per locus which makes a total of 108 possible points to look at. Only 14 of the 108 loci delivered p-values, what leads to 94 loci where the calculation of a p-value was a problem. There were only five out of the 14 loci with p-values smaller than .05. The shown locus is the one with the most significant p-values, which are eight out of the 136 possible comparisons at all. The eighth allele of the eighth locus is used as an example for the analysis, with rounded genotype frequencies p^2, 2pq and q^2 of 0.476, 0.362 and 0.161. There are no missing values at the locus 'L8.08'. Again a bootstrap with 1000 replications is performed to calculate p-values out of the original data for an all pair comparison and letters are used to show the differences between the 17 colonies. The results are shown in Table 9. There are not as much letters used as in the microbov example and there are two more populations. There are three different groups a, b and ab. Only the groups with just a or just b can be separated from each other. The other comparisons show no significant difference.

Table 9: All pair comparison of the 17 cat colonies. Equal letters show that the breeds are not significant different at an α level of .05. The results are from locus 'L8.08'.

Cat colony	Letters of no significance
P02	ab
P03	ab
P04	ab
P05	ab
P06	a
P07	ab
P08	ab
P09	ab
P10	ab
P11	ab
P12	b
P13	a
P14	ab
P15	a
P16	a
P17	b
P01	ab

8 General discussion

The results of the example data sets show that the F_{ST} statistic can be used to differentiate populations into clusters. The simulations show how well the used methods perform. This must be considered at the interpretation of the example data sets. With the used procedure, multiple populations can be separated, depending on the sample size and the difference of the allele frequencies. The power to detect a difference between the population increases with a higher sample size and a greater difference of the allele frequencies. Through the used multiplicity adjustment a global type I error is controlled. Based on the F_{ST} statistic, the bootstrap is used to calculate p-values. The p-value is used to interpret the F_{ST} value. Here it is a data based value to separate populations. Other possible F_{ST} breakpoints could be used if a χ^2 distribution of the test statistics is assumed, which would eventually fit at a high sample size. This is not step with actual practice and so the bootstrap is used to get a breakpoint. There are no breakpoints defined on the null to one scale of F_{ST}, where populations are totally separated or can not be separated, because there are no strict guidelines or recommendations. Therefore the test procedure of this work is used. Obviously the results of significant comparisons depend only somewhat on the number of bootstraps and the population size, if these parameters are already high enough. But the results will be more reproducible the higher the sample sizes and the higher the number of bootstraps are. The effectiveness of more bootstrap replications to receive further information about the F_{ST} distribution depends on the sample size. Waples and Gaggiotti (2006) report a decreasing power of F_{ST} when sample size differs, as it is in the used examples. It is also obvious, that if only one locus is used, the result

depends completely on it and can be different if another one would be reviewed. It is the same with the sampled individuals and populations, because the estimate varies with the sample (Holsinger and Weir, 2009).
At lot of loci from the example data sets, p-values or p-values smaller than .05 could not be estimated. The main reason is probably that there is no difference at these loci. It can also be due to calculation problems with missing values, not available estimates and markers with a low information content. A small sample size of a population can have the same effect which rises in combination with marginal allele frequencies for p^2 or q^2. The smaller the frequency of the heterozygous genotype becomes, the more homozygous genotypes there are, the more discrete the resampled F_{ST} distribution gets and the harder it is to calculate a not too liberal or too conservative p-value. The F_{ST} test statistic is adjusted due to the characteristic that F_{ST} usually can not become negative, to reduce the discreteness.
Todays DNA markers are highly polymorphic, which is the basic requirement of a marker. They are also mostly easily available and can therefore be used to routinely estimate population parameters as F_{ST}. To detect these population structures, without defining exactly what a population is, the used locus has to be representative for population and also for the aim of the work (Waples and Gaggiotti, 2006). F_{ST} has a distinct sensitivity to high polymorphism when migration is low (Hartl and Clark, 1997). Unsampled populations can influence the estimates of parameters as migration rate, genetic diversity and corresponding statistics if this 'ghost' populations interact with the sample (Slatkin, 2005). Whitlock and McCauley (1999) come to the conclusion, that the results of F_{ST} do not apply to the most real populations, because F_{ST} and the variants of it are idealized and have simplistic basics as e.g. the HWE. This is because several points of the HWE are not considered. Evolution requires e.g. mutation and selection so HWE is only an approximation because of the stringent conditions to describe populations at a certain time. Mutation is a very important factor of evolution. There are different types of mutations. Some have influence to the phenotype, some are only observable with molecular markers. Mutations are undirected and happen at different rates, depending of the organism and the region in the genome (Drake et al., 1998). Overall, mutation leads to polymorphism and it can lead therefore to greater differences between populations. This influence depends on migration and selection. Selection works conjoined with mutation, if it results in an advantage or disadvantage in survivability. An example for a disadvantage is, if highly conserved regions mutate. The individuals mostly have a negative survivability, because these changes often lead to death. The mutation model used here is a strong simplification of available models and applicable, because only a single locus and generation are considered. This model can fit with very small genes which have only a few nucleotypes, but these types of genes do not exist. The possibility that a mutation changes the allele A to allele a or a back to A shrinks rapidly if the number of nucleotides increases. For

this case, several models have been developed by scientists. The mutation model e.g. influences the gene frequencies of the immigrant gene pool (Beaumont, 2005). The infinite allele model (IAM, (Kimura and Crow, 1964)), used by F_{ST}, and the stepwise mutation model (SMM, (Kimura and Otha, 1978)) are the most known and there are several other models which are mentioned e.g. in Balloux and Lugon-Moulin (2002). In the IAM each mutation will result in a completely new allele. If homozygous alleles appear, they are identical by descent. Identical by descent in the IAM is an extra form of identity in state. At identity in state only the allelic state of a gene like length or sequence is compared (Balding et al., 2001). A possibility to measure genetic diversity is the probability of sampling two alleles of the same type. In the SMM each mutation changes the number of repeats if a microsatellite marker is used, which includes the allele size into the comparison of individuals or populations. The mutation models are mentioned, because of the different characteristics of molecular markers and because the population differentiation parameters each have one of this model as an assumption. Molecular markers can fit to a mutation model differently, e.g. microsatellite markers appear to be in the middle of the two mutation models IAM and SMM (Balding et al., 2001).

As a result of different mutation rates and selection pressure, the term 'heterogeneous genomic divergence' was created (Nosil et al., 2009). This means that the genome can be highly uneven in terms of genetic differentiation and so the F_{ST} values depend on the used locus. Divergent selection reduces heterogeneity by selecting extreme phenotypes, because that genotypes are often homozygous. Molecular genetic differentiation can happen at specific loci if they are linked to the phenotype, this can act with or without gene flow (Nosil et al., 2009). F_{ST} can also be used to test if loci are influenced by natural selection. This is indirectly done if the Lewontin Krakauer Test is used (Beaumont, 2005). Loci under selection show higher F_{ST} values than neutral loci (Oleksyk et al., 2010). The effect of heterogeneous genomic divergence was observed by Charlesworth (1998). The F_{ST} value depends to big parts on the within population diversity level, which is documented by two DNA variation studies of *Drosophila* where the variation within and between populations was compared. Regions with a low recombination rate did have a higher variability between populations, than genes in regions with a normal recombination rate. Loci with high F_{ST} values did show a reduced within population variability (Charlesworth, 1998).

Besides mutation and selection the genetic drift can change allele frequencies in a population, too. Genetic drift is a complete random change like bottle neck effects of small populations after a mass mortality caused by a catastrophic event. Genetic drift is possible in a short time period where selection and mutation will change the gene pool mostly over multiple generations. If one population is split up because for example of continental drift or human activities, the genetic differences will increase over time. The change of allele frequencies depends, next to mutation and selection, on the population size and on the migration rate. Migration is the exchange of individuals between populations. The

higher the migration rate per generation is, the closer the gene pools of the separated populations become. The migration rate or gene flow can be described as mN_i. F_{ST} can be used to estimate gene flow by $F_{ST} = 1/(4mN + 1)$ (Wright, 1951). Todays approaches can estimate mN_i under migration–drift equilibrium quite well, but are limited to pairwise comparisons under non-equilibrium conditions (Hey et al., 2004) or allow only a simple exponential change in population size and miss changes in population divergence (Palsboll et al., 2007). F_{ST} has a simple relationship of Nm in the IAM but it is unclear how this works with more general models of population structure (Balding et al., 2001). The problem of estimating gene flow and separating populations are effects from unsampled populations, which immigrate into the sampled population (Slatkin, 2005) and that the ideal case in which each sample is a deme is rare. The demes are often mixed, because the boundaries between the samples are often not clear. That leads to an underestimation of the sample variance among populations (Balloux and Lugon-Moulin, 2002). Pearse and Crandall (2004) wrote that estimates as F_{ST} have a greatly reduced informativeness if the *a priori* defined classes of the samples as e.g. 'Bazadais' or colony 'P06' do not represent the actual biological reality. An alternative would be to define populations based on the data only without defining populations e.g. by sampling spots *a priori* but for that F_{ST} can not be used. Genetic diversity between groups can be quantified by F_{ST} but the drawback is the unclear definition of a population and the effect of population sampling. The estimated F_{ST} changes with the definition of a population and obviously with the sampled individuals (Pearse and Crandall, 2004; Waples and Gaggiotti, 2006). A simplification for the situation of population boundaries is the population island model or the isolation by distance or also called neighborhood model. If the migration rates are discontinuous this can be seen as an separated island where each island is an area with random mating and so a deme (Waples and Gaggiotti, 2006; Falconer, 1970; Wright, 1951). Allele frequencies between populations get more equal due to the exchange of genes and the effects of selection and genetic drift are reduced. The fixation of alleles is hardly possible if gene flow between populations is high enough (Balloux and Lugon-Moulin, 2002). Whitlock and McCauley (1999) wrote that $F_{ST} \neq 1/(4Nm + 1)$ because the HWE assumptions are not fulfilled and is to use with caution at migration. There are further known recurrent population effects which are contrary to the assumptions of the HWE. One of them is called 'Wahlund'-effect (Wahlund, 1928). Individuals tend to mate with those that are nearby which is a form of non-random mating. These impacts of local mating will mimic those of inbreeding within a single, well-mixed population. This effect is also mentioned by Palsboll et al. (2007) as 'stepping stone population model'. They describe this behavior like before but expound the correlation between the geographical distance and increasing genetic divergence with gene flow that happens only between nearby populations. Even theoretical completely random mating populations like the European eel (*Anguilla anguilla*) are geographically structured (Balloux and Lugon-Moulin, 2002). The

more distinct the populations are, the less is the similarity. However this effect is called, it leads to more homozygous and less heterozygous allele frequencies.
There are a lot of other methods available to separate populations and estimating mN_i. Suggested by Laurentin (2009) are e.g. next to F_{ST} Nei's G_{ST} (Nei, 1973) and Slatkin's R_{ST} (Slatkin, 1995). They are mostly a further development of Wright's F statistics and look in some points similar to them. Nei (1973) described his G_{ST} as an estimate of F_{ST} for multiple alleles instead of a biallelic locus. G_{ST} is the ratio of the difference of total population heterozygosity and the average subpopulation heterozygosity to the total population heterozygosity. The total population heterozygosity is one, subtracted by the sum of the squared average allele frequencies over subpopulations. The amount of genetic variation for highly variable loci determinates the G_{ST} value, which is not the case for F_{ST}. G_{ST} is a derivation of F_{ST}, because it does not use the allele frequency variance among subpopulations but the expected panmictic heterozygousity (Weicker et al., 2001). Weicker et al. (2001) also reported that F_{ST} and G_{ST} versus Cockerham's θ are correlated with an r^2 of 0.91 by using 39 empirical data sets, but they estimated θ by using F_{ST} and G_{ST} values. G_{ST} does not range between 0 and 1 at all and uses fixed effect sampling (Hedrick, 2005). Fixed effect sampling means, that the characteristics of the sample are consigned to the larger population from where the sample is taken from. Next to the disadvantage, that G_{ST} ignores as F_{ST} the sampling error, (Weicker et al., 2001) G_{ST} was not used, because biallelic loci are considered only. R_{ST} was formulated by Slatkin (1995) and is because of that one of the newest popular population differentiation parameter. R_{ST} is the ratio of the difference of the average squared difference in allele size between all pairs of alleles and the average sum of squares of the differences in allele size within each subpopulations to the average squared difference in allele size between all pairs of alleles. So R_{ST} is estimated by the variances of allele sizes, whereas F_{ST} is estimated by the variances of allele frequencies (Balloux and Lugon-Moulin, 2002). It uses the SMM and the variances of allele sizes for the calculation, instead of the variances of allele frequency and also a different mutation model, instead of the IAM, which are used by F_{ST} (Balloux and Lugon-Moulin, 2002). Disadvantages of R_{ST} are, that it has problems if the mutation model could not be assumed. Another disadvatage of R_{ST} is, that it has a high variance (Balloux and Lugon-Moulin, 2002).
As described, all of the parameters to separate populations have their characteristics, advantages and disadvantages (Kitada et al., 2007; Holsinger and Weir, 2009). Independently of the statistic, events as the effect of unsampled populations will occur and the meanderings of the HWE assumptions and missing values are an issue not only for F_{ST}. The used F_{ST} works at the simulated data and at the example data sets but there are still further improvements which can be programmed, simulated and implemented. This can be simulating different sample sizes, HWE meanderings, missing values or considering multiple loci. At first a less conservative F_{ST} would be the aim. The F_{ST} here is estimated

simply. The variance of the F_{ST} parameter is not estimated. To estimate the variance of the F_{ST} parameter, a bootstrap estimate can be used. This would not be optimal and an extensive procedure. Secondly the implementation of multiple loci to F_{ST} can be done. This is also extensive, because if more than one locus is analysed, the correlation between the loci should be taken into account. Furthermore Charlesworth (1998) mentioned that methods as F_{ST}, which measure the proportion of between-population to total diversity, 'are not necessarily appropriate', if multiple loci with a high variance within populations are compared. So the implementation of G_{ST} or R_{ST} might not be a big step forward, depending on the data. A probable solution would be the implementation of a new population differentiation parameter based on linear models. A further task would be the implementation of a HWE check. Available programs as e.g. Genepop (Rousset, 2008) can calculate this next to F_{ST} or multiple alleles but not pairwise comparisons of populations and multiplicity adjusted p-values.

To sum up it is shown that the used procedure is implementable to separate populations but given the following simplified settings. The populations are compared in a single generation at a single locus. Therefore mutation, selection, migration, linkage disequilibrium between loci etc. can be ignored. The used population sample and the correct classification of the individuals to the populations and their representativeness is assumed. The experimenter has the sole responsibility of the sample quality. Misinterpretation would be the result, if the individuals or compared species do not fit. Results are not interpretable if e.g. independently evolved taxa are compared or more precisely F_{ST} does not work at homoplasy (Balloux and Lugon-Moulin, 2002). But for the used setting 'F_{ST} is an excellent measure of the genetic differentiation among populations' (Whitlock and McCauley, 1999), independent if this difference comes from genetic (e.g. mutation) or statistical sampling variance (e.g. finite sampled individuals). Due to the calculated p-value this measure is easy interpretable and due to the maximal test statistic, which is implemented in this work, the FWER is controlled if multiple comparisons are done.

9 References

Aichinger, E. and Grimm, T. (2008). Hardy-Weinberg equilibrium: first description 100 years ago. *Medizinische Genetik*, 20(3):276–281.

Bachtrog, D., Agis, M., Imhof, M., and C, S. (2000). Microsatellite variability differs between dinucleotide repeat motifs-evidence from *Drosophila melanogaster*. *Molecular Biology and Evolution*, 17:1277–1285.

Balding, D., Bishop, M., and Cannings, C. (2001). *Handbook of Statistical Genetics*. Wiley.

Balloux, F. and Lugon-Moulin, N. (2002). The estimation of population differentiation with microsatellite markers. *Molecular Ecology*, 11(2):155–165.

Beaumont, M. (2005). Adaptation and speciation: what can F_{ST} tell us? *Trends in Ecology and Evolution*, 20(8):435–440.

Büning, H. and Trenkler, G. (1998). *Nichtparametrische statistische Methoden*. de Gruyter Lehrbücher.

Charlesworth, B. (1998). Measures of divergence between populations and the effect of forces that reduce variability. *Molecular Biology and Evolution*, 15(5):538–543.

Davison, A. and Hinkley, D. (1997). *Bootstrap Methods and Their Applications*. Cambridge University Press.

Devillard, S., Jombart, T., and Pontier, D. (2011). Disentangling spatial and genetic structure of stray cat (*Felis catus L.*) colonies in urban habitat using: not all colonies are equal. *submitted to Molecular Ecology*.

Drake, J., Charlesworth, B., Charlesworth, D., and Crow, J. (1998). Rates of spontaneous mutation. *Genetics*, 148(4):1667–1686.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7:1–26.

Efron, B. and Tibshirani, R. (1993). *An Introducion to the Bootstrap*. Chapman & Hall.

Falconer, D. (1970). *Introduction to quantitative genetics*. Ronald Press.

Fisher, R. (1930). *The Genetical Theory of Natural Selection*. Clarendon Press.

Hartl, D. and Clark, A. (1997). *Principles of Population Genetics*. Sinauer Associates, Sunderland, Massachusetts.

Hedrick, P. (2005). A standardized genetic differentiation measure. *Evolution*, 59(8):1633–1638.

Hedrick, P. W. (2000). *Genetics of Populations*. Jones And Bartlett Publishers, 2nd edition.

Hey, J., Won, Y., Sivasundar, A., Nielsen, R., and Markert, J. (2004). Using nuclear haplotypes with microsatellites to study gene flow between recently separated Cichlid species. *Molecular Ecology*, 13(4):909–919.

Holsinger, K. and Weir, B. (2009). Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. *Nature Reviews Genetics*, 10(9):639–650.

Horn, M. and Vollandt, R. (1995). *Multiple Tests und Auswahlverfahren*. Gustav Fischer Verlag.

Jarne, P. and Lagoda, P. (1996). Microsatellites, from molecules to populations and back. *Trends in Ecology & Evolution*, 11(10):424–429.

Kimura, M. and Crow, J. (1964). The number of alleles that can be maintained in a finite populations. *Genetics*, 49:725–738.

Kimura, M. and Otha, T. (1978). Stepwise mutations model and distibution of allelic frequencies in a finite populations. *Proceedings of the National Academy of Sciences of the USA*, 75:2868–2872.

Kitada, S., Kitakado, T., and Kishino, H. (2007). Empirical Bayes inference of pairwise F_{ST} and its distribution in the genome. *Genetics*, 177(2):861–873.

Kumar, P., Gupta, V., Misra, A., Modi, D., and Pandey, B. (2009). Potential of Molecular Markers in Plant Biotechnology. *Plant Omics*, 2(4):141–162.

Laloe, D., Jombart, Tand Dufour, A., and Moazami-Goudarzi, K. (2007). Consensus genetic structuring and typological value of markers using multiple co-inertia analysis. *Genetics Selection Evolution*, 39(5):545–567.

Laurentin, H. (2009). Data analysis for molecular characterization of plant genetic resources. *Genetic Resources and Crop Evolution*, 56(2):277–292.

Leviyang, S. (2010). The distribution of F_{ST} for the island model in the large population, weak mutation limit. *Stochastical analysis and applications*, 28(4):577–601.

Moritz, C. (1994). Defining evolutionarily-significant-units for conservation. *Trends in Ecology & Evolution*, 9(10):373–375.

Myles, S., Davison, D., Barrett, J., Stoneking, M., and Timpson, N. (2008). Worldwide population differentiation at disease-associated SNPs. *BMC Medical Genomics*, 1:1–22.

Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences of the United States of America*, 70(12):3321–3323.

Nosil, P., Funk, D., and Ortiz-Barrientos, D. (2009). Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, 18(3):375–402.

Oleksyk, T. K., Smith, M. W., and O'Brien, S. J. (2010). Genome-wide scans for footprints of natural selection. *Philosophical Transactions of the Royal Society B-Biological Sciences*, 365(1537):185–205.

Palsboll, P., Berube, M., and Allendorf, F. (2007). Identification of management units using population genetic data. *Trends in Ecology & Evolution*, 22(1):11–16.

Pearse, D. and Crandall, K. (2004). Beyond F_{ST}: Analysis of population genetic data for conservation. *Conservation Genetics*, 5(5):585–602.

Piepho, H. (2004). An algorithm for a letter-based representation of all-pairwise comparisons. *Journal of Computational and Graphical Statistics*, 13(2):456–466.

R Development Core Team (2010). R: A language and environment for statistical computing. Vienna, Austria. ISBN 3-900051-07-0, URL: http://www.r-project.org.

Rousset, F. (2008). GENEPOP ' 007: a complete re-implementation of the GENEPOP software for Windows and Linux. *Molecular Ecology Resources*, 8(1):103–106.

Slatkin, M. (1995). A measure of population subdivision based on microsatellite allele frequencies. *Genetics*, 139(1):457–462.

Slatkin, M. (2005). Seeing ghosts: the effect of unsampled populations on migration rates estimated for sampled populations. *Molecular Ecology*, 14(1):67–73.

Wahlund, S. (1928). Zusammensetzung von Population und Korrelationserscheinung vom Standpunkt der Vererbungslehre aus betrachtet. *Hereditas*, 11:65–106.

Waples, R. and Gaggiotti, O. (2006). What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Molecular Ecology*, 15(6):1419–1439.

Weber, J. and Wong, C. (1993). Mutation of human short tandem repeats. *Human Molecular Genetics*, 2(8):1123–1128.

Weicker, J., Brumfield, R., and Winker, K. (2001). Estimating the unbiased estimator theta for population genetic survey data. *Evolution*, 55(12):2601–2605.

Weir, B. and Cockerham, C. (1984). Estimating F-statistics for the analysis of population-structure. *Evolution*, 38(6):1358–1370.

Weir, B. and Hill, W. (2002). Estimating F-statistics. *Annual Review of Genetics*, 36:721–750.

Westfall, P. and Young, S. (1993). *Resampling - Based Multiple Testing*. Wiley-Interscience.

Whitlock, M. and McCauley, D. (1999). Indirect measures of gene flow and migration: F_{ST} not equal 1/(4Nm+1). *Heredity*, 82(Part 2):117–125.

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

Wright, S. (1965). The interpretation of population-structure by F-statistics with special regard to systems of mating. *Evolution*, 19(3):395–420.

A Appendix

A.1 R-code

In the following section the program code for the statistical software R (R Development Core Team, 2010) is shown.

A.1.1 Function to generate populations

```
genosim <- function(p, n){
  pg <- cbind(p^2, 2*p*(1-p), (1-p)^2, n)
  apply(pg, 1, function(x){
    sample(rep(0:2, rmultinom(1, size=x[4], x[1:3])))
  })
}
```

A.1.2 Function to estimate allele frequencies

```
########################
### Dominant marker ###
########################
allelfreqDOM <- function(gpop, pop){
  n <- tapply(gpop,pop,length)
  r <- length(unique(pop))
  ## Estimate allele frequencies
  naa <- tapply(gpop, pop, function(x) sum(x == 2))

  #HWE assumed
  q <- naa/n
  p <- 1-q
  h <- 2*p*q
  out <- list(p=p, h=h, n=n)
  out
}

#########################
### Codominant marker ###
#########################
allelfreq <- function(gpop, pop){
  n <- tapply(gpop,pop,length)
  r <- length(unique(pop))
  ## Estimate allele frequencies
  nAA <- tapply(gpop, pop, function(x) sum(x == 0))
  nAa <- tapply(gpop, pop, function(x) sum(x == 1))
  naa <- tapply(gpop, pop, function(x) sum(x == 2))

  pAA <- nAA/n
  pAa <- nAa/n
  paa <- naa/n
  p <- pAA + pAa/2
  q <- paa + pAa/2
  # proportion heterozygous
  h <- nAa/n
  out <- list(p=p, h=h, n=n)
  out
}
```

A.1.3 Bootstrap function

The bootstrap is done at every created population using the *boot* function from the *boot* R-package in version 1.2-43. The *boot* functions are adapted from (Davison and Hinkley, 1997).

```
#####################################
### Bootstrap for two populations ###
#####################################
library(boot)
Ftest <- function(gpop, R=1000){
  geno <- as.vector(gpop)
  pop <- as.factor(rep(paste("pop",1:ncol(gpop), sep=""), times=n))
  af <- allelfreq(geno, pop)
  oF <- Fstats(af$p, af$h, af$n)[1]
  Fboot <- function(geno, inds, pop, oF){
    afb <- allelfreq(geno[inds], pop)
    bF <- Fstats(afb$p, afb$h, afb$n)[1]
    bF - oF
  }
  fboot <- boot(geno, Fboot, strata=pop, R=R, stype="i", pop=pop, oF=oF)
  bstat <- fboot$t[,1]
  pv <- mean(bstat[!is.nan(bstat)] > oF)
  c(Fst=oF, pvalue=pv)
}

#########################################
### Bootstrap for multiple populations ###
#########################################
#All pair comparison
pairwiseFstats <- function(p, h, n){
  require(multcomp)
  K <- abs(contrMat(p, type="Tukey"))
  apply(K,1,function(x) Fstats(p[x > 0], h[x > 0], n[x > 0]))
}

library(boot)
pairwiseFtest <- function(gpop,pop, R=1000){
  af <- allelfreq(gpop, pop)
  oF <- pairwiseFstats(af$p, af$h, af$n)[1,]
  Fboot <- function(gpop, inds, pop, oF){
    afb <- allelfreq(gpop[inds], pop)
    bF <- pairwiseFstats(afb$p, afb$h, afb$n)[1,]
    bF - oF
  }  fboot <- boot(gpop, Fboot, strata=pop, R=R, stype="i", pop=pop, oF=oF)
  bstat <- apply(fboot$t,1,max)
  reject <- na.omit(t(apply(rbind(bstat),2, function(x) x > oF)))
  data.frame(Fst=oF, pvalue=apply(reject,2,mean))
}
```

A.1.4 F_{ST} function

The F_{ST} function is adapted from (Weir and Cockerham, 1984).

```
Fstats <- function(p, h, n){
  r <- length(p)
```

```
  # Weir & Cockerham (1984)
  nb <- mean(n)
  nc <- (r*nb - sum(n^2)/(r*nb)) / (r-1)
  pb <- sum(n*p)/(r*nb)
  s <- sum(n*(p-pb)^2)/((r-1)*nb)
  hb <- sum(n*h) / (r*nb)
  #
  a <- (nb/nc) * ( s - (1/(nb-1)) * ( pb*(1-pb) - ((r-1)/r)*s - (1/4)*hb))
  b <- (nb/(nb-1)) * ( pb*(1-pb) - ((r-1)/r)*s - ((2*nb-1)/(4*nb))*hb)
  c <- 1/2*hb
  #
  theta <- a / (a+b+c)
  F <- 1 - c / (a+b+c)
  f <- 1 - c / (b+c)
  #
  out <- c(theta, F, f)
  names(out) <- c("theta", "F", "f")
  out
}
```

A.2 Figures of the comparison of two populations

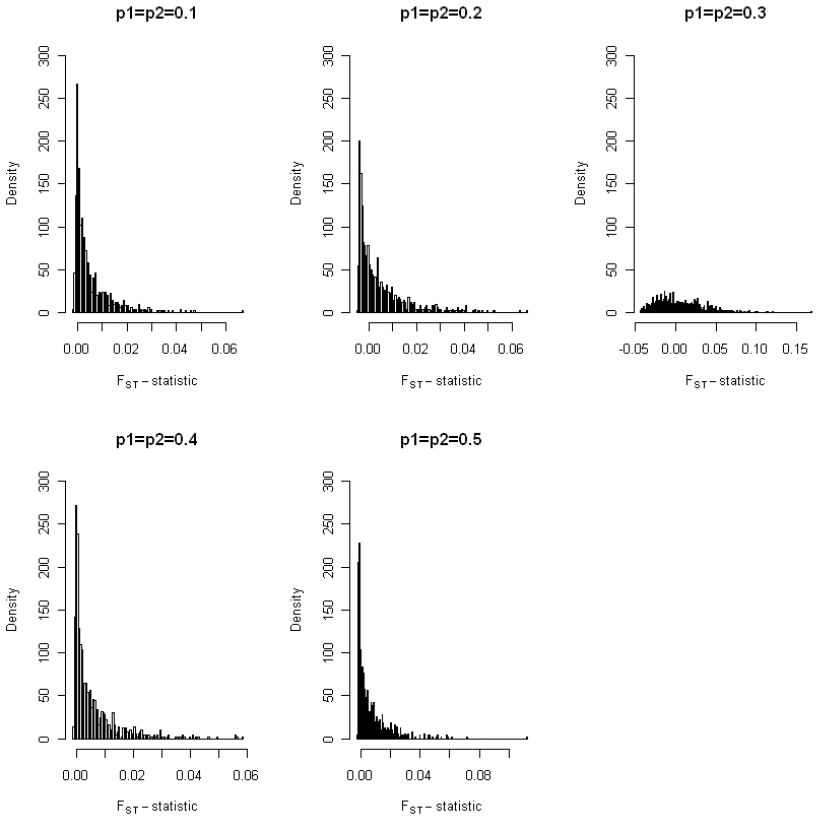

Figure 8: Simulated F_{ST} distributions from 1000 bootstraps for the allele frequency p with 0.01, 0.025, 0.05, 0.1, 0.2, 0.3, 0.4 and 0.5 for both of the two generated populations with 100 individuals. Missing values are set to null.

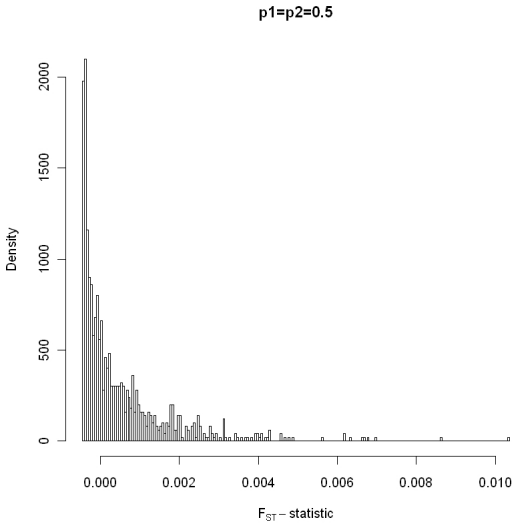

Figure 9: Simulated F_{ST} distribution of two populations with 1000 individuals from 1000 bootstraps for the allele frequency p = 0.5.

A.3 Figures and tables of the global comparison of multiple populations

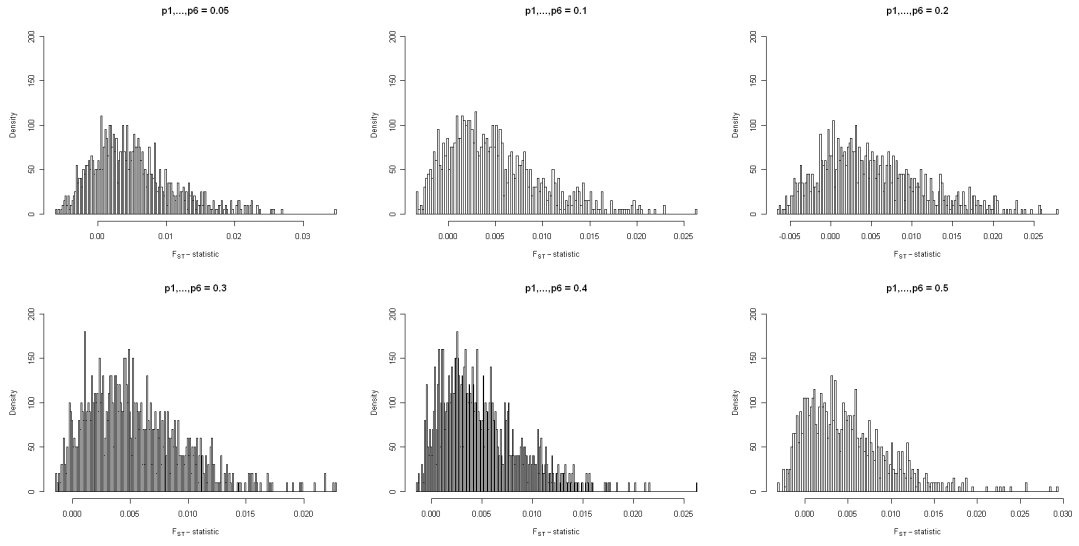

Figure 10: Simulated global F_{ST} distributions of six populations from 1000 bootstraps for the allele frequency p with 0.05, 0.1, 0.2, 0.3, 0.4 and 0.5 for all of the populations with 100 individuals. Missing values are set to null.

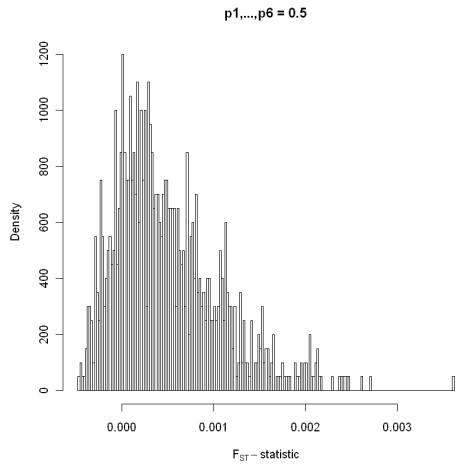

Figure 11: Simulated global F_{ST} distribution of six populations of a global comparison from 1000 bootstraps for the allele frequency p = 0.5, for all of the populations with 1000 individuals.

Table 10: Each of the 14 rows shows one setting of allele frequencies for each of the six populations and the corresponding means of simulated F_{ST} values. The populations are compared with global comparisons. Four different population sizes are tested.

Allele frequencies	n = 5	n = 10	n = 25	n = 50
0.15, 0.15, 0.15, 0.15, 0.15, 0.15	-0.00	-0.00	-0.00	0.00
0.25, 0.25, 0.25, 0.25, 0.25, 0.25	0.00	0.00	-0.00	0.00
0.5, 0.5, 0.5, 0.5, 0.5, 0.5	0.00	-0.00	-0.00	0.00
0.75, 0.75, 0.75, 0.75, 0.75, 0.75	-0.00	-0.00	-0.00	0.00
0.85, 0.85, 0.85, 0.85, 0.85, 0.85	0.00	-0.00	-0.00	0.00
0.5, 0.5, 0.5, 0.5, 0.5, 0.15	0.08	0.08	0.08	0.08
0.5, 0.5, 0.5, 0.5, 0.5, 0.25	0.04	0.04	0.04	0.04
0.5, 0.5, 0.5, 0.5, 0.5, 0.75	0.04	0.04	0.04	0.04
0.5, 0.5, 0.5, 0.5, 0.5, 0.85	0.08	0.08	0.08	0.08
0.1, 0.1, 0.1, 0.1, 0.1, 0.15	0.00	0.00	0.00	0.00
0.1, 0.1, 0.1, 0.1, 0.1, 0.25	0.03	0.03	0.03	0.03
0.1, 0.1, 0.1, 0.1, 0.1, 0.5	0.18	0.18	0.19	0.19
0.1, 0.1, 0.1, 0.1, 0.1, 0.75	0.39	0.40	0.40	0.40
0.1, 0.1, 0.1, 0.1, 0.1, 0.85	0.50	0.49	0.49	0.49

A.4 Figures and tables of the all pair comparison of multiple populations

Table 11: Each of the 14 rows shows one setting of allele frequencies for each of the six populations and the corresponding means of simulated F_{ST} values. The populations were compared with an all pair comparison. Four different population sizes are tested.

Allele frequencies	n = 5	n = 10	n = 25	n = 50
0.15, 0.15, 0.15, 0.15, 0.15, 0.15	-0.01	0.00	0.00	0.00
0.25, 0.25, 0.25, 0.25, 0.25, 0.25	-0.01	0.00	0.00	0.00
0.5, 0.5, 0.5, 0.5, 0.5, 0.5	-0.01	0.00	0.00	0.00
0.75, 0.75, 0.75, 0.75, 0.75, 0.75	-0.01	0.00	0.00	0.00
0.85, 0.85, 0.85, 0.85, 0.85, 0.85	-0.01	0.00	0.00	0.00
0.5, 0.5, 0.5, 0.5, 0.5, 0.15	0.21	0.23	0.24	0.24
0.5, 0.5, 0.5, 0.5, 0.5, 0.25	0.10	0.12	0.12	0.12
0.5, 0.5, 0.5, 0.5, 0.5, 0.75	0.10	0.12	0.12	0.12
0.5, 0.5, 0.5, 0.5, 0.5, 0.85	0.21	0.23	0.24	0.24
0.1, 0.1, 0.1, 0.1, 0.1, 0.15	0.00	0.01	0.01	0.01
0.1, 0.1, 0.1, 0.1, 0.1, 0.25	0.05	0.06	0.07	0.07
0.1, 0.1, 0.1, 0.1, 0.1, 0.5	0.28	0.30	0.31	0.32
0.1, 0.1, 0.1, 0.1, 0.1, 0.75	0.57	0.59	0.60	0.60
0.1, 0.1, 0.1, 0.1, 0.1, 0.85	0.69	0.71	0.71	0.72

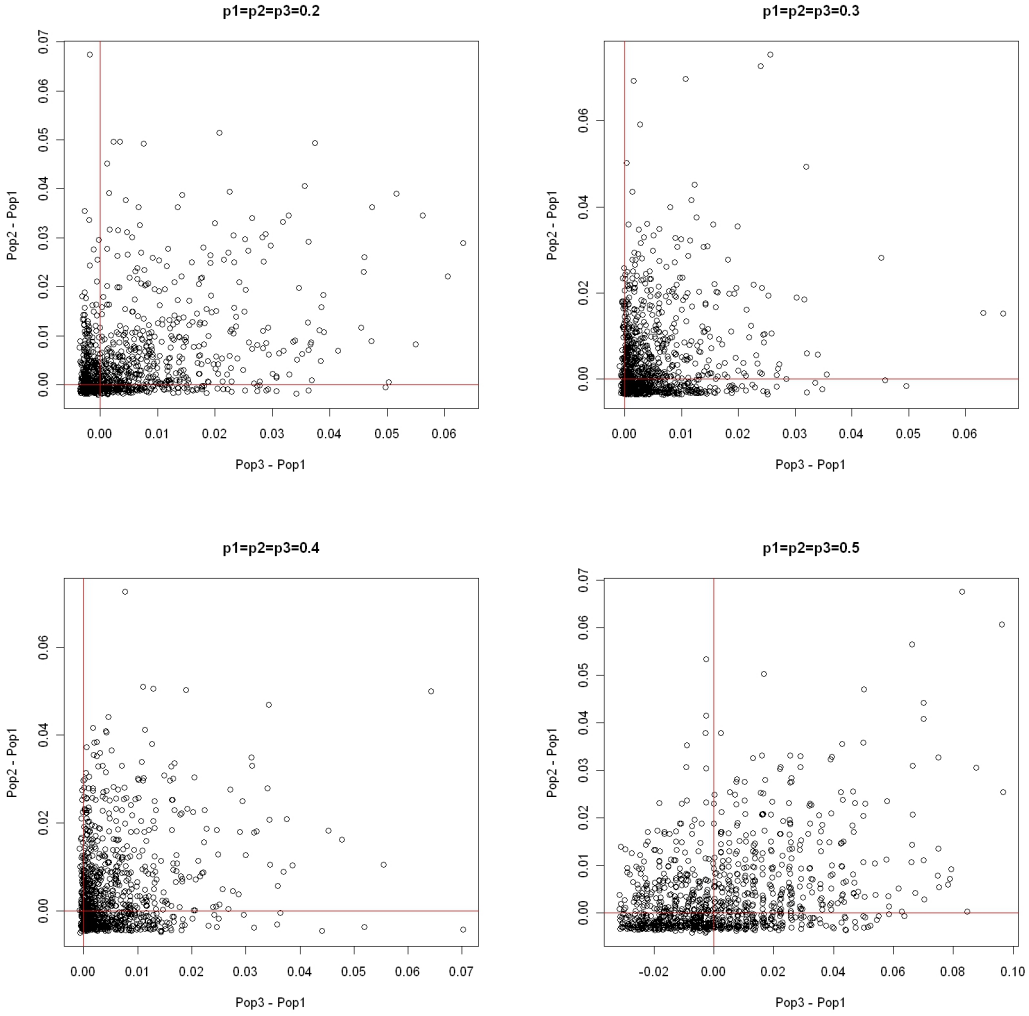

Figure 12: Simulated multidimensional F_{ST} distributions of three populations from 1000 bootstraps for the allele frequency p_i with 0.2, 0.3, 0.4 and 0.5 for all of the populations with 100 individuals. Missing values are set to null.

A.5 Table of α simulations

Table 12: Saturation of α for twelve different population sizes. Either two or six populations with an allele frequency of 0.5 at a single locus are compared. 10000 replications and 1000 bootstraps are done.

Populations size	two populations	six populations (global)	six populations (all pair)
75	1e-04	0	0
100	1e-04	0	0
150	1e-04	0	0
200	0	0	0
250	0	0	0
500	0	0	0
1000	1e-04	0	0
1500	2e-04	0	0
2000	0	0	0
2500	0	1e-04	0
3000	0	0	0
3500	0	0	0

A.6 Additional information of the example data sets

A.6.1 Extract of the microbov example data set

Table 13: Extract of 704 reviewed cattle individuals of 373 loci of 15 populations. The results of the analysis with genetic marker data are shown. The possible genotypes of the biallelic loci are indicated with 0, 0.5 and 1. The required information of the corresponding population of the individuals is in this case not combined with the results of the individual data but saved in an extra slot.

	L01.1	L01.2	L01.3	L01.4	L01.5	L03.02
001	0	0	0	0	0	0
002	0	0	0	0	0	0.5
003	0	0	0	0	0.5	1
004	0	0	0	0	0	0
005	0	0	0	0	0.5	0
006	0	0	0	0	0.5	0
007	0	0	0	0	0.5	0
008	0	0	0	0	0	0
009	0	0	0	0	0.5	0.5
010	0	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	
600	0	0	0	0	1	0
601	0	0	0	0.5	0.5	0
602	0	0	0	0	1	0
603	NA	NA	NA	NA	NA	0
604	0	0	0	0	1	0
605	0	0	0	0.5	0.5	0
606	0	0	0	0	1	0
607	0	0	0	0	0.5	0.5
608	0	0	0	0.5	0.5	0.5
609	0	0	0	1	0	0
610	0	0	0	0.5	0.5	0
611	0	0	0	1	0	NA
612	0	0	0	0	1	0
613	0	0	0	0	1	NA
614	0	0	0	1	0	0
615	0	0	0	0.5	0.5	0
⋮	⋮	⋮	⋮	⋮	⋮	
700	0	0	0	0.5	0	0.5
701	0	0	0	0	0	0
702	0	0	0	0.5	0	0
703	0	0	0	1	0	0
704	0	0	0	0.5	0	1

A.6.2 Multiplicity adjusted and not adjusted p-values of the all pair comparison of the microbov data set

Table 14: Multiplicity adjusted p-values of the all pair comparison of the microbov data set of locus 'L03.02'.

	Bazadais	BlondeAquitaine	Borgou	BretPieNoire	Charolais	Gascon	Lagunaire	Limousin	MaineAnjou	Montbeliard	NDama	Salers	Somba	Zebu	Aubrac
Bazadais	1	0.387	0.874	0.727	0.866	0.924	0.000	1	0.001	0.999	0.098	1	1	0.000	1
BlondeAquitaine		1	1	1	1	1	0.695	0.622	1	1	1	0.695	0.626	0.999	0.892
Borgou			1	1	1	1	0.262	0.971	0.980	1	1	0.981	0.970	0.965	0.998
BretPieNoire				1	1	1	0.212	0.895	0.997	1	1	0.919	0.890	0.994	0.989
Charolais					1	1	0.298	0.965	0.981	1	1	0.978	0.965	0.970	0.997
Gascon						1	0.190	0.988	0.961	1	1	0.991	0.988	0.933	0.999
Lagunaire							1	0.000	1	0.001	0.931	0.000	0.000	1	0.000
Limousin								1	0.018	1	0.216	1	1	0.007	1
MaineAnjou									1	0.507	1	0.025	0.017	1	0.080
Montbeliard										1	0.976	1	1	0.410	1
NDama											1	0.264	0.208	1	0.473
Salers												1	1	0.017	1
Somba													1	0.007	1
Zebu														1	0.056
Aubrac															1

Table 15: Not multiplicity adjusted p-values of the all pair comparison of the microbov data set of locus 'L03.02'.

	Bazadais	BlondeAquitaine	Borgou	BretPieNoire	Charolais	Gascon	Lagunaire	Limousin	MaineAnjou	Montbeliard	NDama	Salers	Somba	Zebu	Aubrac
Bazadais	1	0.018	0.101	0.079	0.080	0.118	0.000	1	0.000	0.250	0.005	1	1	0.000	0.600
BlondeAquitaine		1	0.637	1	0.611	0.508	0.001	0.047	0.204	0.296	0.473	0.030	0.020	0.168	0.065
Borgou			1	1	1	1	0.001	0.144	0.127	0.640	0.319	0.158	0.116	0.088	0.213
BretPieNoire				1	1	1	0.003	0.121	0.249	0.470	0.458	0.111	0.089	0.205	0.167
Charolais					1	1	0.001	0.133	0.097	0.613	0.314	0.134	0.103	0.084	0.194
Gascon						1	0.000	0.144	0.092	0.924	0.282	0.158	0.130	0.083	0.220
Lagunaire							1	0.000	0.192	0.000	0.066	0.000	0.000	0.205	0.000
Limousin								1	0.000	0.342	0.011	1	1	0.000	1
MaineAnjou									1	0.048	0.657	0.000	0.000	1	0.002
Montbeliard										1	0.176	0.342	0.303	0.029	0.443
NDama											1	0.004	0.008	0.556	0.026
Salers												1	1	0.000	1
Somba													1	0.000	1
Zebu														1	0.002
Aubrac															1

A.6.3 Multiplicity adjusted and not adjusted p-values of the all pair comparison of the nancycats data set

Table 16: Multiplicity adjusted p-values of the all pair comparison of the nancycats data set of locus 'L8.08'.

	P02	P03	P04	P05	P06	P07	P08	P09	P10	P11	P12	P13	P14	P15	P16	P17	P01
P02	1	1	1	1	1	1	1	1	1	1	0.292	1	1	1	1	0.535	1
P03		1	1	1	1	1	1	1	1	1	0.418	1	1	1	0.997	0.711	1
P04			1	1	0.959	1	1	1	1	1	0.883	0.978	1	0.995	0.939	0.983	1
P05				1	0.723	1	1	1	1	1	0.979	0.850	1	0.933	0.659	1	1
P06					1	0.573	1	1	1	0.995	0.004	1	1	1	1	0.004	0.488
P07						1	1	1	1	1	0.995	0.681	1	0.821	0.504	1	1
P08							1	1	1	1	0.261	1	1	1	1	0.533	1
P09								1	1	1	0.173	1	1	1	1	0.362	1
P10									1	1	0.150	1	1	1	1	0.317	1
P11										1	0.594	0.998	1	1	0.989	0.824	1
P12											1	0.005	0.161	0.009	0.004	1	0.998
P13												1	1	1	1	0.008	0.617
P14													1	1	1	0.327	1
P15														1	1	0.013	0.768
P16															1	0.004	0.403
P17																1	1
P01																	1

Table 17: Not multiplicity adjusted p-values of the all pair comparison of the nancycats data set of locus 'L8.08'.

	P02	P03	P04	P05	P06	P07	P08	P09	P10	P11	P12	P13	P14	P15	P16	P17	P01
P02	1	1	0.428	0.334	0.284	0.268	1	1	1	0.993	0.001	0.418	1	0.453	0.243	0.020	0.321
P03		1	0.834	0.484	0.245	0.409	1	1	1	1	0.007	0.369	1	0.380	0.206	0.077	0.436
P04			1	1	0.031	0.737	0.573	0.463	0.371	0.998	0.032	0.106	0.392	0.114	0.029	0.135	0.787
P05				1	0.026	1	0.412	0.329	0.274	0.479	0.088	0.057	0.281	0.060	0.015	0.277	1
P06					1	0.003	0.290	0.312	0.325	0.086	0.000	1	0.356	1	1	0.000	0.020
P07						1	0.349	0.294	0.216	0.410	0.151	0.028	0.251	0.042	0.003	0.293	1
P08							1	1	1	0.999	0.003	0.403	1	0.441	0.262	0.047	0.396
P09								1	1	0.887	0.000	0.454	1	0.402	0.253	0.019	0.334
P10									1	0.701	0.000	0.460	1	0.427	0.242	0.006	0.323
P11										1	0.004	0.204	0.595	0.212	0.053	0.039	0.434
P12											1	0.000	0.000	0.000	0.000	1	0.244
P13												1	0.473	1	1	0.000	0.069
P14													1	0.522	0.281	0.008	0.282
P15														1	1	0.000	0.084
P16															1	0.000	0.007
P17																1	0.433
P01																	1

A.6.4 Multiplicity adjusted and not adjusted p-values of the all pair comparison of the fungus data set

Table 18: Multiplicity adjusted p-values of the all pair comparison of the fungus data set of 'Locus4'.

	Dortmund	GGH	Kassel	Kordes	Noack	Ruthe	Sangerhausen	StHannover	Tantau	UNI	UNISORTEN	Bremen
Dortmund	1	0.995	1	0.892	0.454	0.995	0.287	1	1	0.585	1	1
GGH		1	1	1	0.960	0.490	0.006	1	0.995	0.945	1	1
Kassel			1	1	0.937	0.737	0.030	1	1	0.922	1	1
Kordes				1	1	0.082	0.001	1	0.930	0.995	1	0.947
Noack					1	0.001	0.000	0.867	0.494	1	0.787	0.397
Ruthe						1	0.931	0.764	0.993	0.008	0.805	0.985
Sangerhausen							1	0.030	0.228	0.000	0.030	0.145
StHannover								1	1	0.879	1	1
Tantau									1	0.634	1	1
UNI										1	0.794	0.577
UNISORTEN											1	1
Bremen												1

Table 19: Not multiplicity adjusted p-values of the all pair comparison of the fungus data set of 'Locus4'.

	Dortmund	GGH	Kassel	Kordes	Noack	Ruthe	Sangerhausen	StHannover	Tantau	UNI	UNISORTEN	Bremen
Dortmund	1	0.310	0.262	0.063	0.001	0.402	0.021	0.369	1	0.000	0.485	1
GGH		1	1	0.726	0.224	0.191	0.002	1	0.357	0.120	1	0.445
Kassel			1	0.235	0.027	0.196	0.000	1	0.340	0.001	1	0.547
Kordes				1	0.261	0.038	0.000	0.326	0.071	0.058	0.414	0.194
Noack					1	0.000	0.000	0.091	0.002	0.795	0.203	0.032
Ruthe						1	0.330	0.227	0.402	0.000	0.276	0.369
Sangerhausen							1	0.002	0.010	0.000	0.010	0.030
StHannover								1	0.429	0.013	1	0.674
Tantau									1	0.000	0.543	1
UNI										1	0.038	0.003
UNISORTEN											1	0.997
Bremen												1

Eidesstattliche Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe, dass alle Stellen der Arbeit, die wörtlich oder sinngemäß aus anderen Quellen übernommen wurden, als solche kenntlich gemacht sind und dass die Arbeit in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegt wurde.

Hannover, den